Multimodal Product Keyword Generation with Adaptive Semantic ID-Based Augmentation

Anonymous ACL submission

Abstract

Product keyword generation plays a critical role in advertising retrieval systems by expanding product-query matching and mitigating zero-result queries. Existing approaches predominantly rely on product textual descriptions and historical query logs, which often omit fine-grained product attributes and fail to capture diverse user expressions, especially for mid- and low-frequency queries. In this work, we propose a multimodal keyword generation framework that leverages product images to recover missing attributes and introduces semantic ID-based clustering to expand keyword diversity across semantically similar products. Specifically, we first extract multi-attribute keywords from multimodal product information, including text and images. We then use RQ-VAE to generate semantic IDs and apply an adaptive granularity strategy—using attribute-level SIDs for bounded attributes and product-level SIDs for high-dispersion attributes—to cluster products and mine additional attribute-consistent keywords for training data augmentation. The augmented dataset is used to fine-tune a multimodal large language model for product keyword generation. Experiments on a large-scale real-world food delivery dataset show that incorporating multimodal attributes significantly improves keyword coverage, while semantic clustering further enhances keyword diversity and retrieval performance. Our method achieves substantially higher matching rates than recent state-of-the-art approaches and demonstrates competitive results on public benchmarks while requiring a significantly simpler inference pipeline.

1 Introduction

Keyword generation is a core component of advertising retrieval systems, aiming to generate keywords for queries or products to enhance recall coverage and mitigate issues such as missed or zero-result queries(Li et al., 2019; Long et al., 2025). On the query side, traditional approaches typically rely on query rewriting or expansion models to produce variant queries, thereby broadening the scope of relevant product retrieval(Lee et al., 2018; Wang et al., 2024; Lei et al., 2024, 2025; Lian et al., 2019). On the product side, existing methods often leverage product attribute information and online behavioral metrics, such as click-through and conversion rates, to generate descriptive, packaging-oriented keywords that summarize product characteristics(Li et al., 2019; Zhou et al., 2019; Wang et al., 2025b; Chen et al., 2025).

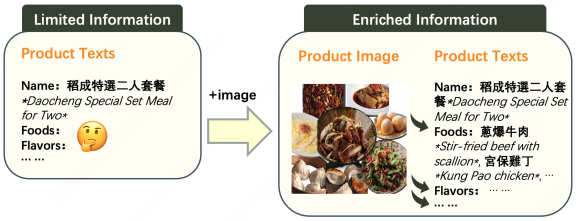

Figure 1: Original product information (**left**) versus product information enhanced with image understanding (**right**). It can be observed that relying solely on textual descriptions fails to capture the detailed contents of product listings. In contrast, incorporating image information enables the extraction of fine-grained content details and additional product attributes, such as foods and flavors.

However, existing product-side methods predominantly rely on textual descriptions and historical click data. OKG (Wang et al., 2025b) and OMS (Chen et al., 2025) leverage product text or high-frequency search terms but remain constrained by incomplete product descriptions. RARE (Liu et al., 2025) proposes a real-time ad retrieval framework that uses a customized LLM to generate Commercial Intent (CI) terms as an intermediate semantic representation and applies constrained beam search to establish query–CI–product mappings. While RARE has shown strong performance in practice, it still relies on product

Attributes	Food	Ingredients	Categories	Flavors	Cooking_methods	Functional_uses	Synonyms	Similar_foods
Examples	葱爆牛肉 *Stir-fried beef with scallion*, 宫保雞丁 *Kung Pao chicken*, ···	牛肉 *Beef*, 三文魚 *Salmon*, 豆腐 *Tofu* ,···	四川菜 *Sichuan cuisine*, 健康餐 *Healthy meals*, ···	麻辣 *Spicy and numbing*, 酸爽 *Sour and refreshing*, ···	爆炒 *Stir-frying*, 乾煎 *Pan-frying*, 水煮 *Boiling*, ···	聚餐分享 *Sharing in group meals*, 減脂瘦身 *Weight loss*, ···	燒烤 *Barbecue*, 烤串 *Skewered grilled meat*, ···	雞胸肉沙律 *Chicken breast salad*, 生菜沙律 *Lettuce salad*, ···

Figure 2: Analysis of click log samples where user queries failed to match product text reveals eight distinct categories of challenging queries. The figure shows the names of the eight categories and representative examples.

text and does not exploit visual modalities, leading to missed recall opportunities for queries involving attributes unmentioned in product descriptions. As illustrated in Figure 1, many product texts omit key details such as foods and flavors that are clearly visible in product images. Furthermore, even when images recover missing attributes, diversity across regional and cultural query variants remains a challenge.

To address these limitations, we propose a representation-driven framework centered on **Semantic IDs (SIDs)**. SIDs define a learned discrete semantic space that organizes products and attributes into transferable neighborhoods, enabling structured, neighborhood-level supervision beyond instance-level annotations. The framework has two key components: (1) leveraging multimodal (text + image) data to recover missing product attributes, and (2) SID-based clustering with **adaptive granularity selection** to mine attribute-consistent keywords from semantically similar products. Concretely, we employ GPT-4.1 (OpenAI et al., 2024) to generate keywords across eight attribute dimensions, train RQ-VAE to assign discrete SIDs at the appropriate granularity—attribute-level for bounded attributes and product-level for high-dispersion attributes—cluster products to form shared keyword pools, and augment the training set via similarity-based keyword selection. Removing the SID component leads to clear performance degradation, confirming the gain stems from structured semantic representation rather than trivial module combination.

The main contributions are:

1. A **large-scale multimodal dataset** constructed from a Hong Kong food delivery platform, where product images and text are jointly processed by GPT-4.1 to generate keywords across eight attribute dimensions, refined via multi-stage LLM-based filtering and human verification to ensure annotation quality.

2. A **representation-driven augmentation framework** centered on Semantic IDs (SIDs) with **adaptive granularity selection**: attribute-level SIDs for bounded attributes and product-level SIDs for high-dispersion attributes, enabling structured neighborhood-level keyword transfer across semantically similar products. Combined with multimodal inputs that recover missing attributes from images, this framework improves both keyword coverage and diversity.

3. **Extensive experiments** demonstrating consistent gains across multiple LLM architectures, significantly outperforming SOTA on query-keyword-product matching, achieving competitive performance on public benchmarks with a simpler inference pipeline, and showing consistent improvements on Amazon-M2 and ImplicitAVE evaluations.

2 Preliminary Study

To investigate the impact of missing product attributes on keyword generation, we conducted preliminary experiments that revealed key challenges in existing methods. We sampled approximately 200,000 high-click-through-rate product listings from the Hong Kong region. Our analysis found that about **40%** of the products lacked detailed descriptions. For example, a listing like "*NVN South Vietnamese Baguette: liver pate, smoked ham, pork sausage, pig's ear, pickled radish, cucumber, mayonnaise, scallion oil, and herbs*" omits ingredients like "smoked ham" or "pickled radish." Without these details, the product is only retrievable by exact search terms like "*NVN South Vietnamese Baguette*", causing missed retrievals for queries with ingredients like "*smoked ham*". This leads to gaps in keyword coverage, preventing users from finding products with specific attributes.

To mitigate this, we incorporated product images alongside titles into GPT-4.1 to extract missing attribute information. This approach recov-

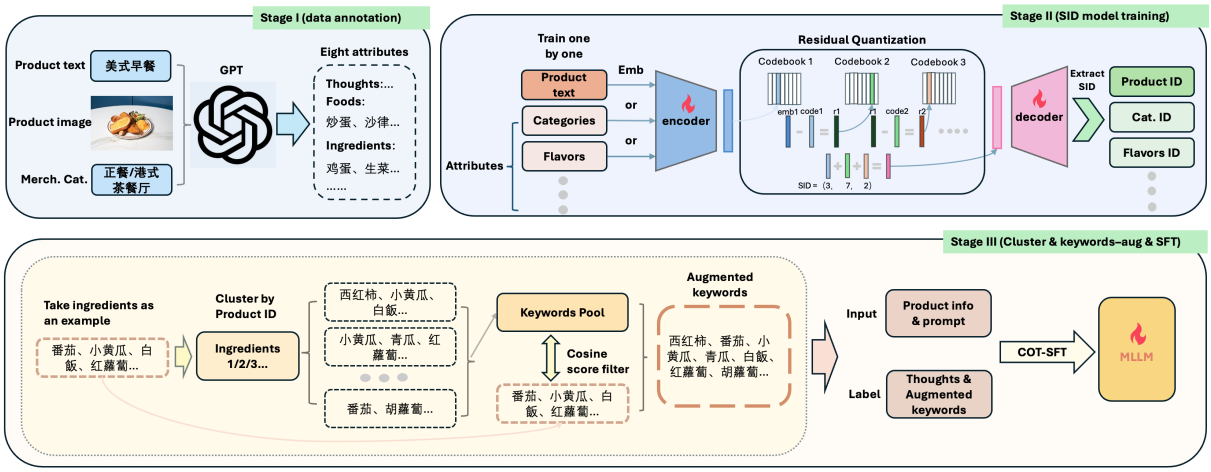

Figure 3: The overall framework of our method comprises three stages. In **Stage I**, multimodal product data (text and images) are fed into GPT-4.1 to generate multi-attribute keywords, which are refined via LLM-based evaluation and human annotation. In **Stage II**, the product text and selected attribute keywords are used to train semantic ID models, assigning semantic IDs to products and keywords. In **Stage III**, SIDs of attribute keywords are clustered, and additional keywords are retrieved from the cluster keyword pool to augment the training data. The augmented dataset is then used to fine-tune a local multimodal LLM for improved keyword generation.

ered approximately **17%** of previously unmatched queries; however, a substantial fraction remained unresolved.

By analyzing the remaining unmatched samples, we identified eight attribute types (Figure 2) where user queries are most prone to retrieval failures. Guided by these insights, we employed a multimodal LLM to generate keywords for each attribute type, underscoring the need for a comprehensive keyword generation framework.

3 Methods

3.1 Dataset Construction

We define eight attribute-specific categories for product keywords: foods, ingredients, categories, flavors, cooking methods, functional uses, synonyms, and similar foods. In Stage I (illustrated in Figure 3), we use GPT-4.1 (OpenAI et al., 2024) to process the product image, product text, and the merchant's main business category (which may differ from the product's category but serves as reference). GPT-4.1 is prompted to generate a detailed description of the image, explain its reasoning, and summarize keywords for the eight attributes.

To reduce noise in the generated keywords, we apply a two-stage cleaning process. First, we prompt GPT-4.1 to re-evaluate and filter out irrelevant or incorrect keywords, ensuring the quality of the generated list. Then we manually inspect each entry to refine the keywords and remove inaccuracies. To ensure diversity in the test set, crucial for evaluating similar-product keyword mining, we conduct multiple sampling iterations to increase label diversity.

Finally, we construct a dataset with **30K** training samples and **5K** test samples for Hong Kong takeaway food products.

3.2 Semantic ID Model Training

We formulate a discrete quantization and reconstruction task to generate semantic ID representations for product texts and their eight types of keywords. A key design principle is **adaptive SID granularity selection**: rather than applying a uniform SID formulation across all attributes, we assign SID representations at the granularity that best matches each attribute's semantic structure.

Concretely, *bounded* attributes—categories, flavors, cooking methods, and functional uses—have compact vocabularies, making attribute-level SIDs effective for tight semantic clusters. *High-dispersion* attributes—food, ingredients, synonyms, and similar foods—exhibit large vocabularies and many-to-many semantic mappings; attribute-level SIDs fragment the space and degrade keyword transfer. For these, product-level SIDs capture richer contextual semantics, yielding more coherent neighborhood structure. This adaptive design reflects the principle that the appropriate granularity of semantic representation should be matched to each attribute's intrinsic properties. Ablation results validating this design are in Section 4.3.

We construct SID representations using RQ-VAE (Lee et al., 2022) (Stage II of Figure 3). RQ-VAE encodes a text document via BERT (Devlin et al., 2019) into embedding x, which is mapped through a DNN encoder to latent z (residual $r_0 = z$). At each quantization level d, a codebook $\mathcal{C}_d = \{e_k\}_{k=1}^{N}$ selects the nearest entry:

$$c_d = \arg\min_k \|r_d - e_k\|_2. \tag{1}$$

The residual is updated as $r_{d+1} = r_d - e_{c_d}$. After D quantization levels, the quantized vectors $(e_{c_0}, \ldots, e_{c_{D-1}})$ are obtained, and their indices $(c_0, \ldots, c_{D-1})$ form the semantic ID (SID). The final quantized latent representation is $\hat{z} = \sum_{d=0}^{D-1} e_{c_d}$.

The decoder reconstructs the input representation from $\hat{z}$. The overall RQ-VAE training objective is

$$L = L_{\text{recon}} + \beta L_{\text{commit}}, \tag{2}$$

where $L_{\text{recon}} = \|x - \hat{x}\|_2^2$ denotes the reconstruction loss and β controls the weight of the commitment term. The commitment loss across quantization levels is

$$L_{\text{commit}} = \sum_{d=0}^{D-1} \|r_d - \text{sg}[e_{c_d}]\|_2^2, \tag{3}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator.

Codebooks are updated via EMA (Morales-Brotons et al., 2024) and initialized from cluster centers of the first batch to mitigate codebook collapse (Rajput et al., 2023). Products sharing the same SID prefix exhibit similar semantics, providing stable boundaries for neighborhood-level keyword mining.

3.3 Semantic Clustering and Keyword Augmentation

To enhance the diversity of keywords in the training set, we augment each product's attribute-specific keywords by mining terms from products with semantically similar attributes.

For each attribute, we first cluster products based on the first m layers of their attribute-specific semantic IDs (SIDs). Let $S = \{s_1, \ldots, s_n\}$ denote the set of attribute keyword sets for all n products, where s_i contains all keywords associated with the i-th product for this attribute. The cluster C_k is defined as:

$$C_k = \{s_j \mid f_m(SID(x_j, attr)) = k,\ s_j \in S\}, \tag{4}$$

where f_m extracts the first m characters of the attribute-specific SID, and products with the same $f_m(SID)$ prefix are assigned to the same cluster C_k.

Next, for each cluster C_k, we construct a **keywords pool** $V(C_k)$ that contains all keywords for the given attribute across products in the cluster. For the current product x, let $W(x)$ denote its existing set of keywords for this attribute. We then compute the cosine similarity $S(w, v)$ between each keyword $w \in W(x)$ and each candidate keyword $v \in V(C_k) \setminus W(x)$.

To select additional keywords, we take the top-a keywords from $V(C_k)$ that have the highest similarity to any keyword in $W(x)$ but are not already included in $W(x)$:

$$A(x) = \underset{v \in V(C_k) \setminus W(x)}{\arg\text{top-}a} \left\{ \max_{w \in W(x)} S(w, v) \right\}. \tag{5}$$

Here, $A(x)$ is the set of augmented keywords for the current product and attribute, which are then added to the product's label in the training set.

This procedure is applied to all attribute dimensions of every product. By leveraging semantic clustering and similarity-based selection, the method expands the attribute-specific keyword sets, improving both the diversity and coverage of the training data. The workflow is illustrated in Stage III of Figure 3.

3.4 Fine-tuning of the Keyword Generation Model

Finally, based on the augmented training set, we perform chain-of-thought supervised fine-tuning (CoT-SFT) on the MLLM and optimize the model via the next-token prediction loss:

$$P(x) = \prod_{t=1}^{T} P(x_t | x_{<t}), \tag{6}$$

During inference, the model already possesses the required multimodal and diverse generation capabilities. By simply providing the product's text, image, and merchant category, we can generate the product's keywords across the eight specified attribute dimensions.

Model	Setting	Macro		Micro	
		Recall	F1	Recall	F1
Closed-source models					
Qwen-VL-Plus	–	33.57	34.21	26.36	29.92
GPT-4o-2024-11-20	–	31.99	37.51	26.67	33.32
Open-source models					
Qwen3-VL-2B-Instruct	SFT	34.78	36.86	28.35	31.77
	aug_data SFT	37.35	36.55	31.30	32.23
Qwen3-VL-4B-Instruct	SFT	36.71	38.94	30.36	34.12
	aug_data SFT	39.18	38.31	32.73	34.02
Qwen3-VL-8B-Instruct	SFT(Text)	34.16	35.08	28.16	30.68
	SFT(Image)	30.54	31.65	25.32	27.55
	SFT(I+T)	37.31	**39.52**	30.71	34.42
	aug_data SFT	**40.13**	39.12	**33.59**	**34.81**

Table 1: Main results on the HK dataset. Closed-source models serve as zero-shot references only. For open-source models, we apply the proposed SID augmentation strategy to the Qwen3-VL series; unimodal and multimodal SFT settings are further compared on **Qwen3-VL-8B-Instruct**. The key controlled comparison is **SFT(I+T)** vs. **aug_data SFT**: same backbone, same I+T inputs, same training procedure—differing only in SID-based data augmentation. Best results in each column are in bold.

Method	Full Match	Partial Match
Text	41.21	59.78
Image	23.85	47.66
I+T	42.71	60.32
aug_data	**44.22**	**61.43**

Table 2: Evaluation of query-product matching effectiveness on offline product click logs. Queries are segmented using Jieba. **Full match** indicates that all query segments appear in the product string, while **partial match** means at least one segment is matched.

Matching Criterion	Precision
Exact match	0.52
Semantic match	0.94

Table 3: Agreement between GPT-4.1 and human annotations on 200 randomly sampled instances. Exact matching requires identical keywords, while semantic matching allows paraphrased but equivalent expressions.

4 Experiments

4.1 Experimental Setup

Datasets. Our dataset is constructed from product click logs in the Hong Kong region of a large food delivery platform, yielding approximately **35K products** (30K training, 5K test) with an average of **36 GPT-4.1-generated keywords per product**, refined via automatic filtering and human verification. As shown in Table 3, GPT-4.1 achieves a strict precision of 0.52 and semantic precision of 0.94 against human annotations. The gap between exact and semantic precision is mainly attributable to paraphrases and stylistic variations (e.g., "crab meat" vs. "crab"), rather than factual errors; the high semantic precision confirms reliable supervision quality. Per-attribute statistics are in Appendix A.7.

For external validation, we also evaluate on **ImplicitAVE** (Zou et al., 2024), a product attribute value extraction benchmark covering multiple e-commerce categories (Food, Clothing, Footwear, Home, Jewelry & Apparel), where performance is measured by accuracy. We further evaluate on the **Amazon-M2** dataset (Jin et al., 2024) to assess SID augmentation in a text-only structured attribute generation setting.

Evaluation Metrics. Keyword generation is evaluated using **Recall** and **F1-score** via exact keyword matching. The main results table (Table 1) reports both **macro-averaged** and **micro-averaged** metrics. Unless otherwise specified, Recall and F1 in text refer to **micro-averaged** values; precision supplementary results are provided in Appendix A.1.

Implementation Details. Semantic IDs are generated with RQ-VAE trained on 768-dim BERT embeddings. The RQ-VAE uses a DNN encoder (512, 256, 128 with ReLU), producing 32-dim latent representations with 4 residual quantization levels and

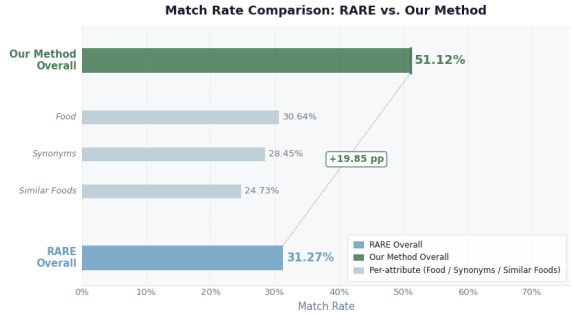

Figure 4: Comparison of the Matching Rate between **RARE Method** and **Our Method** on Click Log Data. Note: The percentages of the three attributes sum to more than the total Match rate because some samples match multiple attributes.

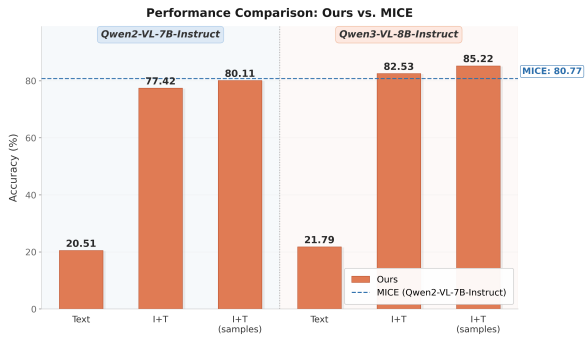

Figure 5: Comparison of **Our Method** against the **MICE Method** on the Food subset of the **ImplicitAVE** benchmark. *"samples"* denotes the addition of three SID-retrieved in-context examples to the prompt.

codebooks of size 256 ($\beta = 0.25$), trained for 500 epochs with AdamW (lr=0.001, batch size 1024). For SFT, we use LoRA with AdamW (lr=2×10^{-4}), cosine annealing, batch size 4, gradient accumulation 8, one epoch on two Tesla A100-80GB GPUs.

Baselines. We evaluate on multiple open-source MLLMs: InternVL3_5-2B/8B (Wang et al., 2025a), Qwen2.5-VL-3B/7B-Instruct (Bai et al., 2025b), and Qwen3-VL-2B/4B/8B-Instruct (Bai et al., 2025a). Closed-source models Qwen-VL-Plus and GPT-4o-2024-11-20 serve as zero-shot reference points. We also compare against **RARE** (Liu et al., 2025), a state-of-the-art commercial intent keyword generation method that applies constrained beam search to establish query-keyword-product mappings. Since the official checkpoint is not publicly available, we re-implement its core pipeline under the same backbone for a fair methodological comparison.

Clustering Method	Recall	F1
k-means	32.45	33.82
SID	**33.59**	**34.81**

Table 4: Ablation study on clustering methods.

Codebook Layers	ID Layer	Recall	F1
3 layers	layer2	31.65	34.50
4 layers	layer2	**33.59**	34.81
	layer3	31.61	**34.97**
5 layers	layer2	33.59	34.55
	layer3	31.42	34.72
	layer4	31.23	34.73

Table 5: Ablation on the number of codebook layers used in the semantic ID model. "layer2/3/4" indicates clustering based on the first 2, 3, or 4 layers of semantic IDs.

4.2 Main Results

Keyword generation results. Table 1 reports performance on the HK dataset; additional model variants are in Appendix A.4. Closed-source models serve as zero-shot references only.

On Qwen3-VL-8B-Instruct, **SFT(Image)** performs worst, as images alone lack explicit semantic descriptions. **SFT(I+T)** outperforms both unimodal settings, confirming that visual signals recover fine-grained attributes (e.g., foods and flavors) absent from text.

The key controlled comparison is **SFT(I+T)** vs. **aug_data SFT**: same backbone, same inputs, same training—differing only in SID-based augmentation. **aug_data SFT** improves Micro-Recall from 30.71 to 33.59 (+2.88), confirming that neighborhood-level augmentation provides more diverse keyword supervision. Consistent gains across all three model scales (2B, 4B, 8B) demonstrate generalizability beyond a single backbone.

Downstream task results. Table 2 reports query-keyword matching rates on a held-out set of **30K** product click logs. Both visual information and SID augmentation significantly improve matching performance. However, since the click logs originate from a text-based recall system, incorporating images provides only limited additional gains.

Comparison with SOTA. We compare our approach against **RARE** (Liu et al., 2025) re-implemented under the same backbone. Our multi-attribute keywords are organized into attribute-

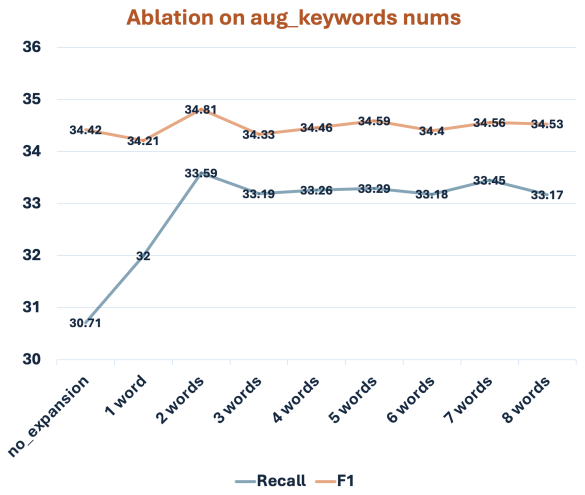

Figure 6: An ablation study on increasing the number of augmented keywords. We compare model **Recall** and **F1** without data augmentation and with augmentation of 1 to 8 words.

Figure 7: Examples of keyword generation. **Green** denotes correct keyword predictions, whereas **red** denotes incorrect keyword predictions.

specific term banks (*food*, *synonyms*, *similar_foods*), and constrained beam search is applied to construct query-keyword-product mappings. As shown in Figure 4, our method achieves a matching rate of **0.5112**, substantially outperforming RARE (**0.3127**).

Results on other benchmarks. As shown in Figure 5, our method achieves competitive performance on the ImplicitAVE Food subset compared with **MICE** (Gong et al., 2025), while using only a single MLLM versus MICE's complex multi-MLLM pipeline involving multiple caption generation and summarization steps. Since each ImplicitAVE sample has a single closed-vocabulary attribute, SFT-based augmentation is not directly applicable; we thus treat this result as supplementary

SID Strategy	Recall	F1
All attribute-level SIDs	32.63	33.58
Adaptive (ours)	**33.59**	**34.81**

Table 6: Ablation on SID granularity strategy (Qwen3-VL-8B-Instruct, aug_data SFT). "All attribute-level SIDs" applies dedicated attribute SIDs to all eight attributes; "Adaptive" uses attribute-level SIDs for bounded attributes and product-level SIDs for high-dispersion attributes.

evidence of the practical usefulness of SID-based keyword generation. We extend evaluation to four additional ImplicitAVE categories (Appendix A.10) and the text-only **Amazon-M2** benchmark (Jin et al., 2024), where SID augmentation improves overall attribute accuracy from **31.87%** to **63.00%** (Appendix A.9), confirming generalizability beyond the food domain.

Joint Multimodal Clustering in RQ-VAE. Fusing CLIP visual features into RQ-VAE clustering slightly improves recall but reduces F1 (Appendix A.12), as visual similarity does not align well with attribute-level semantics—visually similar products may belong to different attribute categories, causing noisy keyword transfer. Text-based SIDs provide more coherent cluster boundaries.

4.3 Ablation Study

Adaptive SID Granularity. A key design choice in our framework is the adaptive SID granularity strategy: attribute-level SIDs for bounded attributes (categories, flavors, cooking methods, functional uses) and product-level SIDs for high-dispersion attributes (food, ingredients, synonyms, similar foods). Table 6 compares this against a uniform attribute-level SID baseline. The adaptive strategy yields consistent gains in both Recall and F1, confirming that SID granularity should be matched to each attribute's intrinsic semantic properties. Per-attribute details are in Appendix A.11.

Clustering Method. To evaluate the advantage of the Semantic ID-based clustering strategy, we compare it with a k-means clustering baseline. For fairness, both methods use the same number of clusters per attribute, and the training set is augmented with up to two keywords per attribute label. As shown in Table 4, SID-based clustering consistently outperforms k-means in both Recall and F1 on Qwen3-VL-8B-Instruct, demonstrating the advantage of semantic ID representations for keyword augmentation.

Number of Codebook Layers and ID Layers. We further analyze the impact of the RQ-VAE codebook depth and the number of ID layers used for clustering. As shown in Table 5, the best balance between Recall and F1 is achieved when `codebook layer=4` and `ID layer=2`. Although setting `codebook layer=5` yields the highest Recall, its lower F1 indicates reduced precision and an inferior overall trade-off.

Number of Augmented Keywords. We also study the effect of the number of augmented keywords in the SID-based keyword augmentation module. As shown in Figure 6, the model achieves the best Recall and F1 when two augmented keywords are used. While increasing the number of augmented keywords improves diversity, excessive augmentation introduces low-similarity noise, leading to performance degradation.

Qualitative Analysis. Figure 7 illustrates typical generation examples. When product texts lack detailed information, text-only models often produce incorrect keywords. In contrast, multimodal inputs and augmented training data help recover missing attributes and reduce mismatches caused by different lexical expressions.

5 Related Work

Product Keyword Generation. Keyword generation improves product-query matching in advertising retrieval. Early approaches used statistical methods (Robertson and Zaragoza, 2009; Ramos, 2003); neural models (Zhou et al., 2019; Li et al., 2019) then leveraged product descriptions for context-aware prediction. Recent LLMs (Liu et al., 2025; Cai et al., 2026) further improve semantic understanding, but most methods rely solely on text, overlooking visual signals. Multimodal models such as GPT-4 (OpenAI et al., 2024), Qwen-VL (Bai et al., 2025b,a), and Gemini (Team et al., 2025) enable fine-grained attribute extraction from images, opening new possibilities for multimodal keyword generation. Nevertheless, recovering missing attributes alone does not ensure keyword diversity, as synonyms and regional variants cannot be inferred from a single product.

Product Semantic Clustering. Grouping similar products is a natural way to address this gap, enabling attribute-consistent keywords to be shared across a neighborhood. Dense representations (e.g., BERT (Devlin et al., 2019)) capture rich product semantics but lack explicit clustering boundaries. Tiger (Rajput et al., 2023) introduces semantic IDs (SIDs) via RQ-VAE (Lee et al., 2022), producing hierarchical discrete codes where similar products share similar ID prefixes, providing a natural mechanism for semantic grouping and keyword mining. We extend this idea by applying SIDs at *adaptive granularity*—attribute-level for bounded-vocabulary attributes and product-level for high-dispersion ones—yielding coherent clusters suited to each attribute type.

6 Conclusion

We presented a multimodal product keyword generation framework that addresses two fundamental challenges in advertising retrieval: missing product attributes and limited keyword diversity. By integrating product images with textual descriptions, our approach recovers critical attribute information missing from text. The core methodological contribution is an **adaptive SID granularity strategy**—attribute-level SIDs for bounded-vocabulary attributes and product-level SIDs for high-dispersion attributes—ensuring that semantic clustering operates at the granularity most suited to each attribute's intrinsic structure. The resulting cluster-based augmentation enriches training data with semantically coherent, cross-product keywords for broader query coverage.

Experiments on our large-scale Hong Kong food delivery dataset demonstrate consistent improvements across multiple model architectures, with augmented fine-tuning significantly outperforming both supervised and zero-shot baselines. Evaluations on Amazon-M2 and ImplicitAVE further confirm the generalizability of the framework. We hope this work highlights the value of representation-driven, multimodal data augmentation for retrieval-oriented NLP tasks.

7 Limitations

Dataset: The current dataset is generated by GPT-4.1, followed by filtering using GPT-4.1 and manual review. However, the keyword generation criteria in the prompt are manually defined. Establishing stricter and more standardized criteria could further improve keyword quality, especially for attributes such as synonyms and similar products, where model performance remains suboptimal. Additionally, since most samples already contain detailed text, the contribution of the image modality may be limited.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Jingwen Cai, Sara Leckner, and Johanna Björklund. 2026. From precision to perception: Human-in-the-loop evaluation of keyword extraction for internet-scale contextual advertising. *Information Systems*, 138:102665.

Bowen Chen, Zhao Wang, and Shingo Takamatsu. 2025. OMS: On-the-fly, multi-objective, self-reflective ad keyword generation via LLM agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18583–18601, Suzhou, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiaying Gong, Hongda Shen, and Janet Jenq. 2025. MICE: Mixture of image captioning experts augmented e-commerce product attribute value extraction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1151–1160, Vienna, Austria. Association for Computational Linguistics.

Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, and 1 others. 2024. Amazon-m2: A multilingual multi-locale shopping session dataset for recommendation and text generation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. ACM.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. *Preprint*, arXiv:2203.01941.

Mu-Chu Lee, Bin Gao, and Ruofei Zhang. 2018. Rare query expansion through generative adversarial networks in search advertising. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 500–508, New York, NY, USA. Association for Computing Machinery.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. 2024. Corpus-steered query expansion with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–401, St. Julian's, Malta. Association for Computational Linguistics.

Yibin Lei, Tao Shen, and Andrew Yates. 2025. ThinkQE: Query expansion via an evolving thinking process. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17772–17781, Suzhou, China. Association for Computational Linguistics.

Zhipeng Li, Jianwei Wu, Lin Sun, and Tao Rong. 2019. Combinatorial keyword recommendations for sponsored search with deep reinforcement learning. *Preprint*, arXiv:1907.08686.

Yijiang Lian, Zhijie Chen, Jinlong Hu, Kefeng Zhang, Chunwei Yan, Muchenxuan Tong, Wenying Han, Hanju Guan, Ying Li, Ying Cao, Yang Yu, Zhigang Li, Xiaochun Liu, and Yue Wang. 2019. An end-to-end generative retrieval method for sponsored search engine –decoding efficiently into a closed target domain. *Preprint*, arXiv:1902.00592.

Tongtong Liu, Zhaohui Wang, Meiyue Qin, Zenghui Lu, Xudong Chen, Yuekui Yang, and Peng Shu. 2025. Real-time ad retrieval via LLM-generative commercial intention for sponsored search advertising. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28948–28960, Suzhou, China. Association for Computational Linguistics.

Hou-Wan Long, Yicheng Song, Zidong Wang, and Tianshu Sun. 2025. Kp-agent: Keyword pruning in sponsored search advertising via llm-powered contextual bandits. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 4991–4995, New York, NY, USA. Association for Computing Machinery.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. 2024. Exponential moving average of weights in deep learning: Dynamics and benefits. *Preprint*, arXiv:2411.18704.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. 2023. Recommender systems with generative retrieval. *Preprint*, arXiv:2305.05065.

Juan Enrique Ramos. 2003. Using tf-idf to determine word relevance in document queries.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Gemini Team, Rohan Anil, and Sebastian Borgeaud. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Yang Wang, Zheyi Sha, Kunhai Lin, Chaobing Feng, Kunhong Zhu, Lipeng Wang, Xuewu Jiao, Fei Huang, Chao Ye, Dengwu He, Zhi Guo, Shuanglong Li, and Lin Liu. 2024. One-step reach: Llm-based keyword generation for sponsored search advertising. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1604–1608, New York, NY, USA. Association for Computing Machinery.

Zhao Wang, Briti Gangopadhyay, Mengjie Zhao, and Shingo Takamatsu. 2025b. OKG: On-the-fly keyword generation in sponsored search advertising. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 115–127, Abu Dhabi, UAE. Association for Computational Linguistics.

Di Wu, Da Yin, and Kai-Wei Chang. 2024. KPEval: Towards fine-grained semantic-based keyphrase evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1959–1981, Bangkok, Thailand. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Yishun Mao, Changlei Zhu, Peng Shu, and Xiaoyan Zhu. 2019. Domain-constrained advertising keyword generation. In *The World Wide Web Conference*, WWW '19, page 2448–2459, New York, NY, USA. Association for Computing Machinery.

Henry Peng Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip S. Yu, and Cornelia Caragea. 2024. ImplicitAVE: An open-source dataset and multimodal LLMs benchmark for implicit attribute value extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 338–354, Bangkok, Thailand. Association for Computational Linguistics.

	Food	Ingredients	Categories	Flavors	Cooking_methods	Functional_uses	Synonyms	Similar_foods
Qwen-VL-Plus	27.62	36.95	56.78	39.78	45.16	32.34	18.07	11.88
InternVL3_5-2B(zero-shot)	13.24	21.5	46.43	13.18	25.1	13.1	2.6	0.84
InternVL3_5-2B(SFT)	27.41	30.3	74.74	27.05	33.26	20.24	8.7	3.15
InternVL3_5-2B(aug_data SFT)	27.29	33.36	73.81	29.18	34.43	21.08	10.45	3.52
InternVL3_5-8B(zero-shot)	20.36	29.25	51.22	22.13	33.24	16.39	7.9	3.61
InternVL3_5-8B(SFT)	32.48	36.67	76.98	31.59	37.76	22.84	11.75	6.09
InternVL3_5-8B(aug_data SFT)	32.87	39.44	77.45	32.47	40.37	23.54	14	6.5
Qwen2.5-VL-3B(zero-shot)	13.64	20.54	54.56	16.25	27.68	10.71	4.9	1.5
Qwen2.5-VL-3B(SFT)	34.07	40.41	83.11	36.8	40.93	22.12	14.07	8.23
Qwen2.5-VL-3B(aug_data SFT)	36.04	45.32	82.53	41.05	43.68	24.78	17.14	8.22
Qwen2.5-VL-7B(zero-shot)	10.41	17.43	54.56	12.6	26.08	11.75	3.67	1.7
Qwen2.5-VL-7B(SFT)	37.41	43.25	82.19	37.27	43.49	24.2	15.43	9.45
Qwen2.5-VL-7B(aug_data SFT)	39.54	47.83	80.82	43.22	44.93	26.59	18.8	10.09
Qwen3-VL-2B(zero-shot)	10.95	17.76	37.36	13.99	17.67	5.68	5.9	1.27
Qwen3-VL-2B(SFT)	38.1	42.62	79.18	35.02	40.12	21.47	14.28	7.43
Qwen3-VL-2B(aug_data SFT)	40.59	47.58	80.8	38.76	41.42	23.77	17.63	8.22
Qwen3-VL-4B(zero-shot)	20.7	31.31	62.27	27.43	35.53	19.67	12.62	4.9
Qwen3-VL-4B(SFT)	41.53	45.16	78.67	37.1	42.22	23.1	16.1	9.84
Qwen3-VL-4B(aug_data SFT)	43.26	49.46	79.56	41.82	43.61	25.3	19.94	10.45
Qwen3-VL-8B(zero-shot)	22.84	32.89	50.66	31.03	39.43	23.92	13.01	6.81
Qwen3-VL-8B(SFT/text)	37.63	41.13	80.75	37.18	42.33	22.78	13.66	8.9
Qwen3-VL-8B(SFT/image)	39.38	34.42	76.41	36.39	41.96	21.98	14.96	8.89
Qwen3-VL-8B(SFT/text+mage	41.76	45.4	81.08	35.73	43.51	22.66	16.6	9.95
Qwen3-VL-8B(aug_data SFT)	45.15	51.01	80.54	41.79	44.94	24.83	21.24	11.57

Figure 8: Evaluation Results of **Recall** for Eight Attribute Keywords under Various Models and Experimental Settings.

	Food	Ingredients	Categories	Flavors	Cooking_methods	Functional_uses	Synonyms	Similar_foods
Qwen-VL-Plus	33.56	43.43	52.32	36.28	41.01	30.59	21.52	14.96
InternVL3_5-2B(zero-shot)	17.64	22.61	45.6	14.13	25.41	13.78	3.33	1.21
InternVL3_5-2B(SFT)	27.32	30.96	68.77	27.52	33.76	22.03	10.56	4.22
InternVL3_5-2B(aug_data SFT)	22.23	28.83	63.86	25.59	31.77	21.65	10.24	4.17
InternVL3_5-8B(zero-shot)	28.03	32.32	47.94	23.41	35.22	17.49	10.23	4.99
InternVL3_5-8B(SFT)	32.16	37.93	71.21	32.06	38.63	24.62	14.12	8.07
InternVL3_5-8B(aug_data SFT)	26.67	34.53	67.22	28.1	37.15	23.52	13.4	7.58
Qwen2.5-VL-3B(zero-shot)	20.32	22.98	53.66	18.7	29.37	12.95	6.39	2.24
Qwen2.5-VL-3B(SFT)	35.28	44.12	79.07	39.78	45.69	25.31	18.29	11.4
Qwen2.5-VL-3B(aug_data SFT)	34.35	44.79	77.2	38.29	45.36	25.91	18.64	10.35
Qwen2.5-VL-7B(zero-shot)	15.5	20.36	51.78	14.63	29.71	13.56	4.87	2.4
Qwen2.5-VL-7B(SFT)	38.29	46.92	77.98	40.22	48.14	27.73	20.02	13.02
Qwen2.5-VL-7B(aug_data SFT)	37.22	46.97	75.66	40.2	47.41	27.77	20.65	12.75
Qwen3-VL-2B(zero-shot)	15.96	19.38	33.32	13.92	17.57	5.57	6.48	1.69
Qwen3-VL-2B(SFT)	38.37	45.84	75.37	37.59	44.38	24.54	18.48	10.3
Qwen3-VL-2B(aug_data SFT)	36.6	45.92	75.28	36.03	43.31	25.3	19.53	10.4
Qwen3-VL-4B(zero-shot)	28.15	35.47	57.87	27.11	37.51	19.32	15.05	6.7
Qwen3-VL-4B(SFT)	41.4	48.56	74.84	39.54	46.65	26.19	20.86	13.46
Qwen3-VL-4B(aug_data SFT)	38.67	47.45	74.44	38.29	45.37	27.53	21.66	13.05
Qwen3-VL-8B(zero-shot)	30.61	38.1	48.18	32.38	41.17	24.04	14.9	8.81
Qwen3-VL-8B(SFT/text)	38.57	44.71	76.28	38.32	45.75	25.31	17.45	12.14
Qwen3-VL-8B(SFT/image)	40.39	37.13	72.34	39.2	46.27	25.13	19.44	12.18
Qwen3-VL-8B(SFT/text+mage)	41.77	48.71	77.06	40.09	47.71	25.78	21.39	13.64
Qwen3-VL-8B(aug_data SFT)	41.26	49.56	74.58	39.05	45.69	27.16	22.6	13.91

Figure 9: Evaluation Results of **F1** for Eight Attribute Keywords under Various Models and Experimental Settings.

A Appendix

A.1 Precision Supplementary Results

Table 7 reports micro-averaged Recall and Precision for key model settings on the HK dataset, complementing the results in Table 1 of the main paper. As discussed in Section 4.3, SID-based augmentation introduces a recall–precision trade-off: augmented training naturally increases the diversity of generated keywords (improving recall) but may also include lower-confidence candidates (slightly reducing precision). The consistent overall improvement confirms that the additional keywords are largely attribute-consistent rather than noisy.

A.2 Main Results by Attribute

Figures 8 and 9 present attribute-level performance across all model variants. After fine-tuning on the augmented training set, Qwen3-VL-8B-Instruct achieves the best or near-best results on most

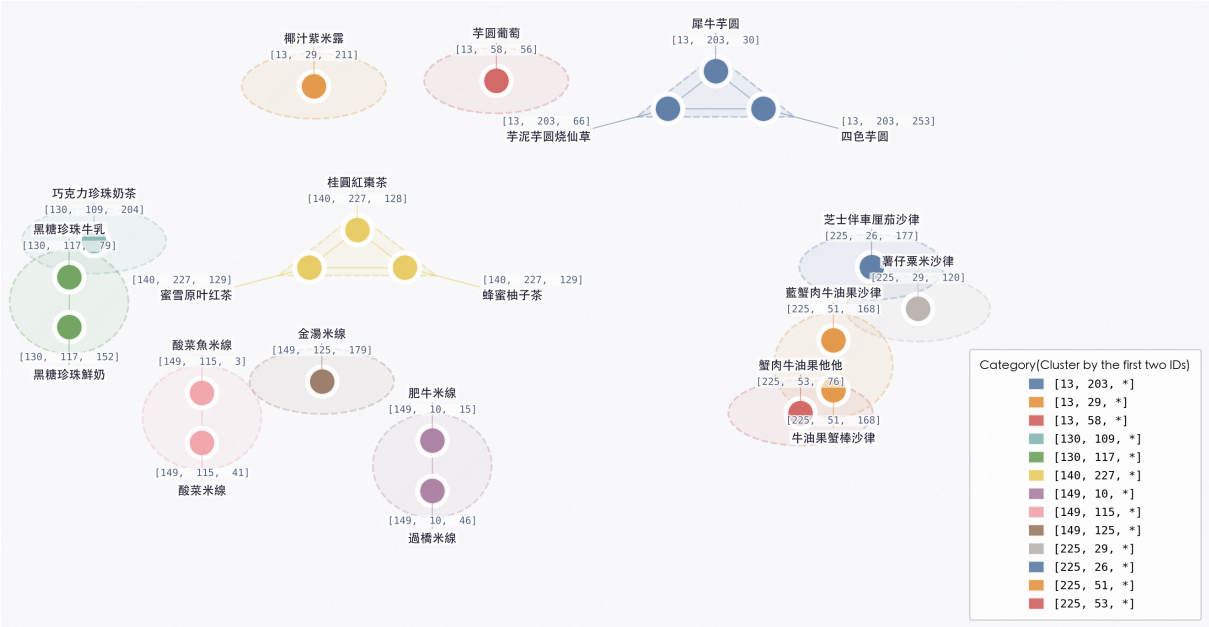

Figure 10: Visualization of selected product text SIDs using Multidimensional Scaling (MDS).

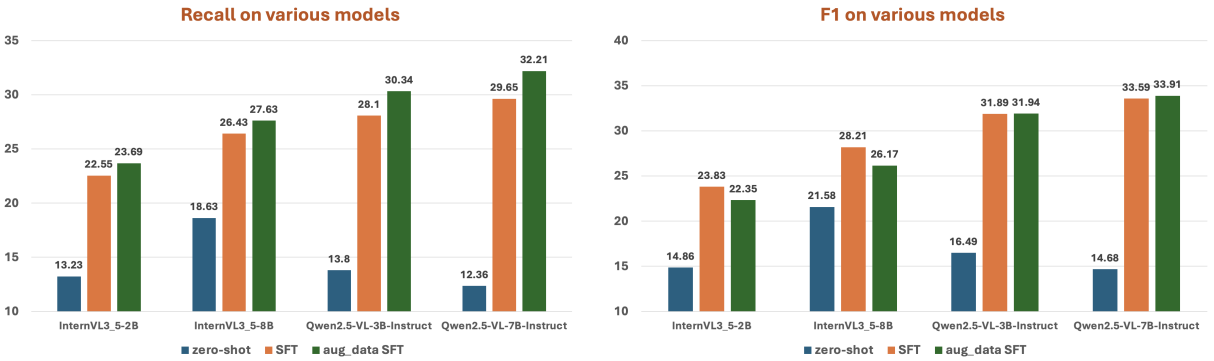

Figure 11: The performance of our method evaluated on **InternVL3_5** series and **Qwen2.5-VL** model series.

Model	Setting	Micro-R	Micro-P
Qwen3-VL-8B	SFT(Text)	28.16	38.69
	SFT(I+T)	30.71	**45.30**
	aug_data SFT	**33.59**	41.08
Qwen3-VL-4B	SFT	30.36	45.11
	aug_data SFT	32.73	40.27
Qwen3-VL-2B	SFT	28.35	41.80
	aug_data SFT	31.30	38.18

Table 7: Micro-averaged Recall (Micro-R) and Precision (Micro-P) for selected settings on the HK dataset. Recall values are reproduced from Table 1; precision results are reported here for completeness.

Codebook Layers	ID Layer	Recall	F1
3 layers	layer2	29.41	32.07
4 layers	layer2	**31.03**	32.23
	layer3	29.16	**32.25**
5 layers	layer2	30.93	31.97
	layer3	29.25	32.27
	layer4	29.04	32.34

Table 8: Ablation study on the number of code layers used in the semantic ID model (**Qwen3-VL-2B-Instruct**).

attribute categories. An exception is the *functional_uses* attribute, where performance is notably lower than that of Qwen-VL-Plus and the gains from data augmentation are limited. This may be due to the inherent difficulty of modeling functional-use semantics and the relatively limited training data available for this attribute.

Codebook Layers	ID Layer	Recall	F1
3 layers	layer2	30.86	33.63
4 layers	layer2	**32.73**	34.02
	layer3	30.84	34.07
5 layers	layer2	32.67	33.71
	layer3	30.73	33.97
	layer4	30.64	**34.15**

Table 9: Ablation study on the number of code layers used in the semantic ID model (**Qwen3-VL-4B-Instruct**).

Attribute	P	R	F1	S-P	S-R	S-F1
food	0.557	0.633	0.579	0.924	0.950	0.936
ingredients	0.622	0.690	0.648	0.939	0.952	0.945
flavors	0.473	0.597	0.512	0.928	0.955	0.941
categories	0.628	0.636	0.628	0.980	0.975	0.978
cooking	0.603	0.744	0.637	0.933	0.961	0.946
func. uses	0.477	0.510	0.482	0.922	0.927	0.924
synonyms	0.351	0.420	0.352	0.941	0.952	0.946
sim. foods	0.419	0.565	0.460	0.917	0.934	0.925
macro avg	**0.516**	**0.599**	**0.540**	**0.936**	**0.951**	**0.942**

Table 10: GPT-4.1 vs. human annotation on 200 samples. P/R/F1: exact match; S-P/S-R/S-F1: semantic match (Wu et al., 2024).

Data Size	Recall	F1
5K	30.11	31.45
10K	31.22	32.61
30K	**33.59**	**34.81**

Table 11: Ablation study on training data size (Qwen3-VL-8B-Instruct, aug_data SFT).

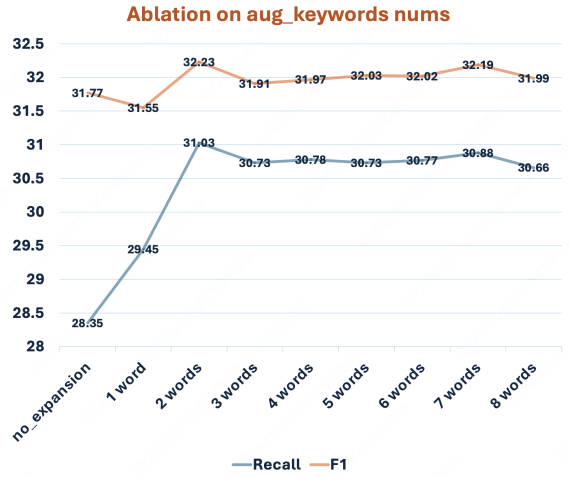

Figure 12: An ablation study on increasing the number of augmented keywords on the **Qwen3-VL-2B-Instruct** model.

A.3 Qualitative Analysis of RQ-VAE SIDs

Figure 10 visualizes selected product text SIDs using multidimensional scaling. Products with similar semantic IDs appear close to each other in the projection and correspond to semantically related items, such as *"Blue Crab Meat and Avocado Salad"* and *"Avocado and Crab Stick Salad"*. In contrast, products with dissimilar SIDs are positioned farther apart; for example, *taro ball*, *salad*, and *rice noodle* products form clearly separated regions. These observations indicate that the learned semantic IDs effectively capture product-level semantic similarity, providing a reliable basis for clustering.

A.4 Main Results on Other Models

We further evaluate our method on the InternVL3_5 and Qwen2.5-VL model series, as shown in Figure 11. The results demonstrate consistent performance improvements across different architectures and model scales, indicating strong robustness of the proposed framework.

A.5 Scaling Law

Table 11 analyzes the impact of training data size by evaluating models trained on 5K, 10K, and 30K samples. Together with the main results, the findings show that model performance consistently improves with increased training data, suggesting that the proposed framework follows a clear scaling trend with respect to both data volume and model capacity.

A.6 Ablation Results on Other Models

Tables 8 and 9 present ablation studies on the number of codebook layers in the semantic ID model and the number of ID layers used for clustering, conducted on Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct. Both models exhibit the same optimal configuration as Qwen3-VL-8B-Instruct, achieving the best balance between Recall and F1 when `codebook_layer=4` and `ID_layer=2`.

Figures 12 and 13 further analyze the number of augmented keywords used in the keyword enhancement module. The optimal configuration is consistent across models: using two augmented keywords yields the best results in both Recall and F1.

Overall, these results indicate that the selected hyperparameters generalize well across different model sizes rather than being specific to a particular architecture.

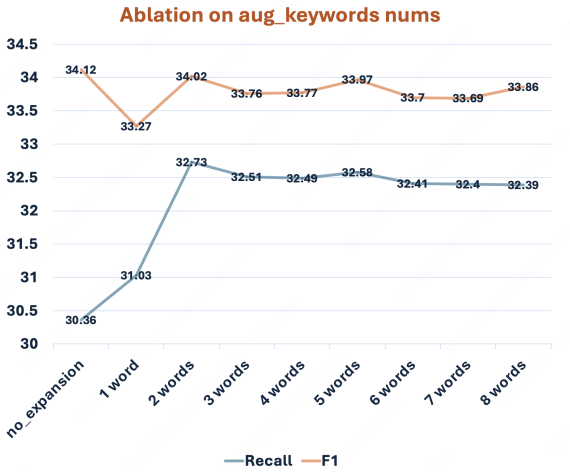

Figure 13: An ablation study on increasing the number of augmented keywords on the **Qwen3-VL-4B-Instruct** model.

A.7 Data Quality Verification

To assess the reliability of GPT-4.1-generated annotations, we randomly sampled 200 products and asked human annotators to independently label the keywords for each attribute. We compared GPT-4.1-generated labels with human annotations.

Table 10 reports both exact-match metrics (Precision, Recall, F1) and semantic-aware metrics (Sem-Precision, Sem-Recall, Sem-F1) (Wu et al., 2024). While exact lexical matching yields relatively lower scores, we observe strong semantic agreement between GPT-4.1 labels and human annotations across all attributes.

Manual inspection of mismatched cases shows that most discrepancies arise from paraphrases or stylistic differences. For example, ingredient descriptions may appear as "crab meat" versus "crab", or dish variations such as "avocado crab salad" versus "crab and avocado salad". Such variations are semantically equivalent but penalized by strict lexical matching.

The semantic-aware metrics, which account for synonymy and paraphrasing, therefore provide a more appropriate evaluation for keyword generation. The high semantic scores indicate that GPT-4.1 labels are largely consistent with human annotations at the meaning level.

Overall, these results suggest that GPT-4.1-generated labels provide reliable semantic supervision and can serve as a practical reference standard for training.

Attribute	SFT (w/o aug)	SFT (w/ SID aug)
Brand	59.82	66.69
Color	31.13	60.19
Size	19.06	60.21
Material	17.46	64.90
Overall	**31.87**	**63.00**

Table 12: Per-attribute accuracy (%) on the **Amazon-M2** dataset. SID-based augmentation is applied in a text-only setting, evaluating structured attribute generation across *brand*, *color*, *size*, and *material*.

Category	Text	I+T	I+T ICL (samples)
Clothing	9.91	79.73	**81.08**
Footwear	38.61	68.67	**70.89**
Home	26.44	71.95	**73.56**
Jewelry & Apparel	25.49	88.24	**90.20**

Table 13: Accuracy (%) on ImplicitAVE categories beyond the Food subset. No additional SFT is performed; all settings use the model without task-specific fine-tuning (see Appendix A.10). **Text**: text-only zero-shot inference. **I+T**: multimodal zero-shot inference. **I+T ICL (samples)**: I+T with three SID-retrieved in-context examples.

A.8 Implementation Details

Experimental Setup for Comparison with RARE. To reproduce the RARE baseline, we use Qwen3-VL-8B-Instruct to generate Commercial Intent (CI) terms from both product images and textual descriptions, producing 25 CIs per product. During constrained beam search decoding, we set `beam_size=50`, `max_step=12`, and the target output length to 25 words. A `length_penalty` of 1.2 is applied to reduce the bias toward shorter outputs.

For our method, we construct attribute-specific keyword banks and select the three most relevant attributes—*food*, *synonyms*, and *similar_foods*—for constrained beam search. Based on the average keyword counts of these attributes, the target output lengths are set to 7, 7, and 11, respectively (25 words in total). The final matching rate is computed as the joint query-attribute-product matching rate across these attributes.

Experimental Setup on the ImplicitAVE Food Subset. Since each sample in the ImplicitAVE dataset contains only a single attribute with one value, the label augmentation strategy used in our main experiments cannot be directly applied. Instead, after incorporating multimodal inputs and generating Semantic IDs (SIDs) for clustering, we

Attribute	Attr-SID	Adaptive
food	43.32	**45.15**
ingredients	49.61	**51.01**
flavors	**81.53**	80.54
categories	41.09	**41.79**
cooking_methods	**45.31**	44.94
functional_uses	24.27	**24.83**
synonyms	20.25	**21.24**
similar_foods	9.92	**11.57**

Table 14: Per-attribute Recall comparing uniform attribute-level SIDs against the adaptive strategy. **Attr-SID**: attribute-level SID applied to all eight attributes. **Adaptive**: our design using attribute-level SIDs for bounded attributes (categories, flavors, cooking methods, functional uses) and product-level SIDs for high-dispersion attributes (food, ingredients, synonyms, similar foods).

RQ-VAE Input	Recall	F1
Text only (ours)	33.59	**34.81**
Text + Image (CLIP)	**33.99**	32.97

Table 15: Overall effect of joint multimodal clustering in RQ-VAE on the HK dataset (Qwen3-VL-8B-Instruct, aug_data SFT).

Attribute	Recall		F1	
	Text	Text+Img	Text	Text+Img
food	**45.15**	43.91	**41.26**	37.48
ingredients	**51.01**	49.11	**49.56**	46.32
flavors	80.54	**87.73**	**74.58**	70.68
categories	41.79	**43.50**	**39.05**	34.43
cooking_methods	44.94	**48.63**	**45.69**	41.05
functional_uses	24.83	**31.40**	27.16	**28.96**
synonyms	**21.24**	20.29	**22.60**	21.74
similar_foods	**11.57**	10.32	**13.91**	12.79

Table 16: Per-attribute Recall and F1 comparing text-only vs. text+image RQ-VAE clustering.

modify the inference procedure accordingly.

Specifically, we augment the prompt with three semantically similar training samples (including their ground-truth labels) retrieved through SID-based clustering. We also append the complete attribute vocabulary to the prompt, constraining the model to generate outputs from this predefined set. No additional training is performed; the model directly performs zero-shot inference on the test set, and performance is evaluated using accuracy.

During SID model training for this dataset, we retain the four-layer codebook structure. Due to the smaller dataset size (approximately 3K products), the codebook size of each layer is reduced to 64, while all other hyperparameters remain consistent with the main experiments.

A.9 Amazon-M2 Per-Attribute Results

Table 12 reports per-attribute accuracy on the Amazon-M2 dataset with and without SID-based augmentation. The text-only SFT baseline achieves an overall accuracy of 31.87%, while SID augmentation raises this to 63.00%. Consistent gains are observed across all four attributes, with the most pronounced improvements on *size* (+41.15 pp) and *material* (+47.44 pp), where label sparsity benefits most from neighborhood-level keyword transfer.

A.10 ImplicitAVE Cross-Category Results

Table 13 reports accuracy on additional ImplicitAVE product categories beyond the Food subset. Following the same protocol as the Food subset, SFT-based augmentation is not applied here for two reasons: (1) each ImplicitAVE sample specifies only a single closed-vocabulary attribute value, making our multi-attribute SFT pipeline incompatible; and (2) the SOTA baseline (MICE) does not use SFT either, so maintaining the ICL setting ensures a fair methodological comparison. We therefore report three settings: **Text** (text-only zero-shot inference), **I+T** (multimodal zero-shot inference), and **I+T ICL (samples)** (I+T augmented with three SID-retrieved in-context examples). SID-based retrieval consistently improves performance across all categories, confirming the generalizability of SID-based retrieval beyond the food domain.

A.11 Adaptive SID Granularity: Per-Attribute Results

Table 14 provides per-attribute Recall comparing the uniform attribute-level SID baseline against our adaptive strategy (overall results are reported in Table 6 of the main paper). The adaptive strategy benefits most from the high-dispersion attributes (food, ingredients, synonyms, similar foods), where product-level SIDs provide more coherent neighborhood structure. For bounded attributes (flavors, cooking methods), attribute-level SIDs remain competitive or slightly better, consistent with their compact vocabularies.

A.12 Joint Multimodal Clustering in RQ-VAE

We investigate whether incorporating visual features into the RQ-VAE clustering stage can further improve keyword augmentation. Specifically, we

fuse CLIP-encoded image features with BERT text embeddings and train RQ-VAE on the joint multimodal representations, keeping all other hyperparameters identical to the text-only baseline.

Table 15 reports the overall results, and Table 16 provides per-attribute breakdown. Multimodal clustering slightly improves recall but reduces F1, indicating lower precision. We attribute this to the gap between visual similarity and attribute-level semantic consistency: visual features expand product neighborhoods by appearance, but may group visually similar yet attribute-different products (e.g., different food categories with similar plating or packaging), causing noisy keyword transfer. This does not mean visual information is unimportant; images remain essential for recovering missing attributes in Stage I. Rather, the result suggests that naive off-the-shelf visual fusion in RQ-VAE is insufficient without task-specific cross-modal alignment, and text-based SIDs provide more semantically coherent cluster boundaries for attribute-level keyword transfer.