Snap2Review: Vision-Grounded Retrieval and Pairwise Preference Alignment for Personalized Reviews Generation

Hongyan Xu[1,2], **Lei Zou**[1], **Dexiang Zhao**[1] and **Xin Fan**[1] **and Chuanwen Luo**[1*]

[1]The School of Information Science and Technology, Beijing Forestry University, Beijing, China
[2]Hebei Key Laboratory of Smart National Park, Beijing Forestry University, Beijing, China
{hongyanxu, zoulei418, zdx7250341, fanxin, chuanwenluo}@bjfu.edu.cn

Abstract

Personalized review generation is vital for e-commerce engagement, yet integrating user-provided images while maintaining distinct user personas remains a critical challenge. Existing methods often struggle to bridge the semantic gap between objective visual signals and subjective linguistic patterns, frequently resulting in hallucinations or homogenized, impersonal content. To address these limitations, we propose Snap2Review, a novel framework that harmonizes cross-modal retrieval with fine-grained preference learning. First, we optimize a cross-modal retriever that maps visual features directly to semantically related historical reviews, using these interactions as precise anchors to strictly ground the generation. Second, we introduce Tri-DPO, which employs a stratified negative sampling strategy with varying difficulty levels, ranging from obvious errors to subtle stylistic mismatches, to force the model to discern fine-grained user preferences. By distinguishing these fine-grained nuances, our model moves beyond generic praise to accurately mimic the user's authentic writing style. Extensive experiments on real-world datasets demonstrate that Snap2Review significantly outperforms strong baselines in both relevance and personalization metrics.

1 Introduction

In the era of visual-centric e-commerce, users increasingly rely on images to narrate their consumption experiences, transforming product reviews from simple text descriptions into rich, multimodal stories. Automatically generating personalized reviews from these user-captured photos holds immense potential to enhance user engagement and streamline feedback loops. However, this task transcends the boundaries of traditional image captioning; it demands an AI agent that acts not merely as an observer describing visual content, but as a personalized ghostwriter capable of articulating the user's unique voice, preferences, and linguistic habits.

Despite the rapid advancements in Multimodal Large Language Models (MLLMs), bridging the chasm between objective visual signals and subjective user personas remains a formidable challenge. Current approaches typically fall into the "Generic Trap": they prioritize factual correctness at the expense of stylistic authenticity. While Retrieval-Augmented Generation (RAG) has been employed to introduce context, standard methods often rely on text-to-text retrieval, failing to capture the subtle visual nuances—such as texture, color deviations, or specific defects—that trigger a user's commentary. Consequently, generated reviews often suffer from a dual failure: Visual Hallucination, where the model invents details unsupported by the image, and Stylistic Homogenization, where the output reads like a robotic, "one-size-fits-all" description rather than a genuine user reflection.

We argue that the root cause of these limitations lies in the coarseness of existing alignment paradigms. Personalization is not a binary objective; a review can be factually correct yet stylistically alien to the user. Standard alignment methods, such as Direct Preference Optimization (DPO) [Rafailov *et al.*, 2023], typically utilize binary preference pairs (winner vs. loser), which are insufficient for teaching a model the fine-grained distinction between a "generic good review" and a "personalized authentic review." To truly mirror a user's persona, the model must learn to navigate a complex preference landscape, distinguishing between semantic errors, sentiment mismatches, and subtle phrasing discrepancies.

To address these challenges, we introduce Snap2Review, a visually-driven framework that harmonizes cross-modal grounding with fine-grained stylistic alignment. Our approach decomposes the personalization problem into two distinct yet interconnected stages.

First, Visual-Anchored Retrieval. To eliminate visual hallucinations, we optimize a Vision-Conditioned Retriever via contrastive learning on image-review pairs. Unlike standard retrievers, this module explicitly aligns visual embeddings with the semantic space of historical reviews, allowing it to retrieve precise textual evidence (e.g., specific mentions of product features) strictly grounded in the input image. This retrieved context serves as a reliable anchor, bridging the gap between raw pixels and semantic understanding.

Second, Fine-Grained Stylistic Alignment via Tri-DPO. To break free from stylistic homogenization, we propose a Persona-Aware Generator trained with our novel Tri-Level

*Corresponding authors.

Direct Preference Optimization (Tri-DPO). Moving beyond binary preference pairs, Tri-DPO employs a stratified negative sampling strategy that constructs negative examples at three distinct difficulty levels: general negatives (irrelevant content), similar negatives (stylistically mismatched but factually correct), and near Negatives (minor semantic or factual errors). This "curriculum of errors" forces the model to discern fine-grained boundaries in user preferences, effectively steering generation from generic praise to authentic, persona-driven expression.

Extensive experiments on real-world e-commerce datasets demonstrate that Snap2Review achieves state-of-the-art performance. By explicitly addressing the twin challenges of visual fidelity and stylistic nuance, our framework generates reviews that are not only relevant to the provided images but also indistinguishable from the user's historical writings in terms of tone and preference.

In summary, the primary contributions of this work are threefold:

- We propose Snap2Review, a unified framework that bridges the modality gap between visual perception and personalized text generation, explicitly addressing the limitations of generic MLLM outputs.
- We introduce a specialized Cross-Modal Retrieval mechanism that grounds generation in visually relevant historical context, effectively mitigating hallucinations caused by purely text-based retrieval.
- We design Tri-DPO, a novel alignment algorithm equipped with a Stratified Negative Sampling strategy. By learning from negative examples of varying difficulty, the model achieves superior alignment with fine-grained user preferences and writing styles.

2 Related Work

2.1 Personalized Review Generation

Automated review generation aims to assist users in articulating experiences and enhancing system interpretability [Li *et al.*, 2020; Peng *et al.*, 2025]. Early works focused on text-based generation, utilizing discrete attributes (e.g., user IDs, ratings) [Tang *et al.*, 2016; Dong *et al.*, 2017] or specific aspect-sentiment labels [Zang and Wan, 2017; Li *et al.*, 2019] to guide RNN or Transformer decoders. While recent approaches leverage historical reviews to improve semantic coherence [Ni and McAuley, 2018; Kim *et al.*, 2020], they largely operate within a unimodal textual space. In modern visual-centric e-commerce, existing Multimodal LLMs often struggle to translate user images into personalized feedback, typically yielding generic content that fails to capture individual writing styles [Truong and Lauw, 2019].

2.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm to mitigate the hallucination issues of LLMs by grounding generation in external knowledge [Fan *et al.*, 2024]. In the context of personalization, RAG has been effectively adapted to retrieve relevant user history or item profiles to guide the model. For instance, [Zhang *et al.*, 2025] utilized a transformer-based retriever to filter essential user preferences, while other works have focused on dynamic prompt selection mechanisms [Huang *et al.*, 2024]. However, most existing RAG frameworks are restricted to text-to-text retrieval, failing to bridge the semantic gap between user-uploaded images and textual history. Standard retrievers overlook critical visual nuances (e.g., texture, defects) essential for accurate reviews. To address this, we optimize a cross-modal retriever that maps visual features directly to historical text, ensuring the generation is strictly anchored in visual reality.

2.3 Preference Alignment in LLMs

Aligning LLMs with human intent is critical for ensuring generation quality [Yu *et al.*, 2025; Huang *et al.*, 2026]. While Reinforcement Learning from Human Feedback (RLHF) laid the foundation, recent research has rapidly expanded alignment techniques to diverse domains.

In specific applications, [Han *et al.*, 2025] employed reinforced self-training to align robot planners with human preferences, while [Yang *et al.*, 2024] focused on behavior alignment in conversational recommendation systems. Techniques have also evolved beyond simple feedback loops; for instance, [Trivedi *et al.*, 2025] introduced a principled approach to prompt optimization for alignment, and [Shankar *et al.*, 2024] investigated the alignment of LLM-assisted evaluation itself. Moreover, in retrieval-augmented scenarios, [Dong *et al.*, 2025] proposed a dual preference alignment framework to better understand LLM needs within RAG pipelines.

Despite these advancements, most approaches rely on binary preference signals or task-specific rewards. For personalized review generation, a binary distinction is insufficient to capture the nuance between semantic correctness and stylistic authenticity. Our work addresses this by introducing a stratified alignment strategy to discern fine-grained stylistic differences.

3 Proposed Method

The proposed Snap2Review framework operates through three progressive stages to achieve personalized, visually-grounded generation. **MR** (Multimodal Retriever Fine-tuning): Aligns visual and textual feature spaces via contrastive learning for cross-modal retrieval; **VGR** (Vision-grounded RAG): Retrieves context to guide the LLM via supervised fine-tuning. **TDO** (Tri-DPO Optimization): Refines stylistic alignment using hierarchically structured negative samples. The overall framework is shown in the Figure 1. The pipeline begins by establishing a robust cross-modal retrieval mechanism, which provides precise semantic anchors for the initial generation phase. Finally, the generated outputs are rigorously polished through our novel preference optimization strategy to ensure deep alignment with the user's specific persona.

3.1 Multimodal Retriever Fine-tuning

To bridge the semantic gap between visual signals and textual descriptions, we employ a dual-encoder architecture that

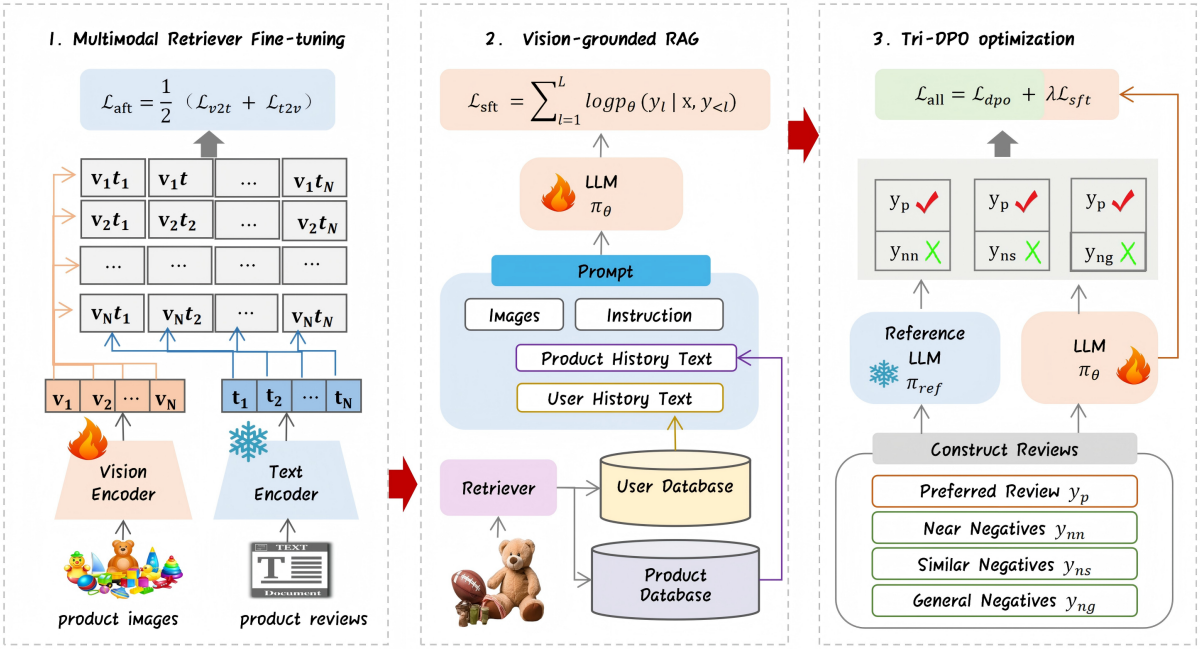


Figure 1: Overview of the Snap2Review framework. The pipeline begins by fine-tuning a cross-modal retriever to support vision-grounded retrieval-augmented generation (RAG), enabling the LLM to utilize relevant historical context. Subsequently, the model is optimized via Tri-DPO using hierarchical negative samples to achieve deep stylistic alignment and visual fidelity.

maps product images and historical reviews into a shared embedding space. This fine-tuned retriever serves as the cornerstone of our framework, ensuring that subsequent generation is grounded in vision-aligned textual evidence rather than generic context.

Domain-Specific Data Collection

We construct a domain-specific dataset $\mathcal{D} = \{(v_i, t_i)\}_{i=1}^{N}$ comprising N pairs of product images and their corresponding user reviews from historical interaction logs. Unlike general-purpose multimodal datasets, these pairs encapsulate e-commerce-specific visual attributes (e.g., texture, defects) and their linguistic descriptors. Training on this corpus enables the model to learn fine-grained visual-textual associations essential for the retrieval task.

Asymmetric Parameter Update

We initialize both the vision encoder f_v and text encoder f_t with pre-trained CLIP weights. To balance domain adaptation with generalization, we adopt an asymmetric fine-tuning strategy.

As shown in the left part of Figure 1, we freeze the parameters of the text encoder f_t to preserve its robust linguistic capabilities while fully fine-tuning the vision encoder f_v. This design choice is motivated by the observation that visual features in e-commerce (e.g., specific product angles, lighting) deviate significantly from general pre-training data, whereas linguistic patterns remain relatively stable. Updating only the vision branch efficiently aligns the visual embeddings to the fixed textual semantic space, reducing computational costs and mitigating overfitting.

Contrastive Learning Objective

We optimize the retriever using a symmetric InfoNCE loss to maximize the similarity between matched image-review pairs while minimizing it for unmatched ones within a batch. Given a batch of B pairs, let $\mathbf{v}_i$ and $\mathbf{t}_i$ denote the normalized embeddings of the i-th image and review, respectively. The similarity score is computed as the dot product $\mathbf{v}_i \cdot \mathbf{t}_i$. The total loss $\mathcal{L}_{aft}$ combines image-to-text ($\mathcal{L}_{v2t}$) and text-to-image ($\mathcal{L}_{t2v}$) alignment objectives:

$$\mathcal{L}_{aft} = \frac{1}{2}(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}), \tag{1}$$

where $\mathcal{L}_{v2t}$ aims to match each product image to its corresponding review. For a batch of B image-review pairs, it is formulated as:

$$\mathcal{L}_{v2t} = -\frac{1}{B}\Sigma_{i=1}^{B} log \frac{exp(\mathbf{v}_i \cdot \mathbf{t}_i/\tau)}{\sum_{j=1}^{N} exp(\mathbf{v}_i \cdot \mathbf{t}_j/\tau)}, \tag{2}$$

where τ is a learnable temperature parameter scaling the logits. The text-to-image loss $\mathcal{L}_{t2v}$ is defined symmetrically. This objective forces the vision encoder to produce embeddings that are semantically proximate to their corresponding textual descriptions, enabling accurate retrieval of historical reviews based on visual queries.

3.2 Vision-Grounded Retrieval-Augmented Generation

Based on the fine-tuned vision encoder, we design a retrieval-augmented pipeline to ground generation in both user history and product context. This process addresses a key limitation of standard LLMs, which often produce generic content,

by grounding the generation in visually relevant and user-specific historical evidence. The framework is shown in the middle part of Figure 1, with three phases: knowledge base construction, joint retrieval, and Persona-Aware Generation.

Hybrid Knowledge Base Construction

We construct two specialized vector databases-a User Database and a Product Database-to capture personal writing styles and item-specific attributes, respectively. To ensure robust matching, we index historical data using a multimodal fusion strategy. For each historical entry (v_i, t_i), we generate visual embedding $\mathbf{V}_i$ and textual embedding $\mathbf{T}_i$ using our fine-tuned retriever. A unified index embedding $\mathbf{e}_i$ is computed as the normalized element-wise sum:

$$\mathbf{e}_i = Normalize(\mathbf{v}_i + \mathbf{t}_i). \tag{3}$$

By fusing modalities, $\mathbf{E}_i$ encodes both the visual appearance of the product and its linguistic description, creating a semantic-rich index that supports accurate retrieval. We index all historical reviews under this unified schema, segregating them by User ID to capture idiosyncratic stylistic patterns and by Product ID to encompass item-specific attributes. This dual-repository architecture provides a comprehensive knowledge base, enabling the system to retrieve evidence that grounds the generation in both the user's personal voice and the product's factual reality.

Joint Retrieval Operation

During inference, given a user-uploaded query image v_q, the system first encodes it into a query vector $\mathbf{v}_q$ using the fine-tuned vision encoder. We then perform a joint similarity search across both databases. The relevance between the query image and a historical entry is measured via cosine similarity:

$$Sim(\mathbf{v}_q, \mathbf{e}_i) = \frac{\mathbf{v}_q \cdot \mathbf{e}_i}{||\mathbf{v}_q|| \times ||\mathbf{e}_i||} \tag{4}$$

We retrieve the top-K entries with the highest similarity scores from each database, yielding a total of $2K$ textual references. This dual-source retrieval ensures the subsequent generation is informed by both the user's historical preferences (Style) and the product's visual attributes (Content).

Persona-aware Generation

The final stage synthesizes the retrieved evidence into a personalized review using a Large Language Model (LLM). This process involves context-aware prompt construction followed by supervised fine-tuning.

Prompt Construction. We structure the input prompt to provide the LLM with comprehensive contextual guidance. The prompt x is composed of four segments:

- **Instruction:** A system-level directive defining the task of personalized review generation.
- **Query Image:** The visual features of the user-provided image v_q.
- **Product Context:** The top-K reviews retrieved from the Product Database, providing factual grounding.
- **User Context:** The top-K reviews from the User Database, serving as stylistic references.

This structured prompt acts as a "persona blueprint," enabling the model to align its output with the user's voice while maintaining factual relevance to the image.

Supervised Fine-tuning (SFT). We first optimize the LLM (π_θ) using standard supervised fine-tuning to adapt it to the generation format. The objective is to maximize the likelihood of the ground-truth review y given the constructed prompt x:

$$\mathbf{L}_{sft} = \sum_{l=1}^{L} logp_\theta(y_l|x, y_{<l}), \tag{5}$$

where L is the length of the target review sequence. This stage equips the model with the foundational capability to integrate visual cues and retrieved context into coherent text, setting the stage for the subsequent preference optimization.

3.3 Tri-Level Negative Preference Alignment

Standard preference alignment methods (e.g., DPO) typically rely on binary pairs (winner vs. loser), treating all non-preferred responses uniformly. However, in personalized review generation, a response can fail in multiple ways: it might be visually hallucinated, stylistically mismatched, or completely irrelevant. To capture these fine-grained distinctions, we propose a Tri-Level Negative Sampling Strategy.

Hierarchical Negative Sampling

As illustrated in the right part of Figure 1, we construct a curriculum of negative samples (y_{nn}, y_{ns}, y_{ng}) for each preferred response y_p. This hierarchy forces the model to learn boundaries at varying levels of difficulty:

- Near Negatives (y_{nn}, Hard): Constructed by applying minor factual/attribute corruptions to the preferred review (e.g., swapping a color/size/material attribute). These are the most challenging samples, teaching the model to learn fine-grained decision boundaries.
- Similar Negatives (y_{ns}, Medium): Sampled from reviews of the *same product* but written by *different users*. These provide correct visual context but incorrect user personas, forcing the model to distinguish between "what to say" and "how to say it."
- General Negatives (y_{ng}, Easy): Sampled from unrelated products and users. These serve as obvious baselines for relevance and coherence.

By contrasting y_p against this spectrum of negatives, the model progressively refines its understanding from coarse-grained relevance (vs. y_{ng}) to fine-grained personalization (vs. y_{nn}).

Tri-DPO Optimization Objective

We formulate the alignment objective to simultaneously penalize all three levels of negatives while maximizing the margin for the preferred response. Let π_θ be the policy model and π_{ref} be the reference model. We extend the standard DPO loss to a multi-negative setting:

$$\mathcal{L}_{dpo} = -\mathbb{E}[log\sigma(\beta log\frac{\pi_\theta(y_p|x)}{\pi_{ref}(y_p|x)} - \tag{6}$$

$$\frac{1}{3}\sum_{i\in nn,ns,ng} \beta log\frac{\pi_\theta(y_i|x)}{\pi_{ref}(y_i|x)})]. \tag{7}$$

Dataset	Method	ROUGE-1			ROUGE-2			ROUGE-L			BertScore
		P	R	F1	P	R	F1	P	R	F1	
Toys	Qwen	16.86	31.27	25.02	3.34	3.87	3.05	15.01	18.57	14.13	52.55
	Gemma	26.03	37.27	25.55	4.50	7.48	4.55	13.24	20.73	13.14	50.41
	InternVL	26.86	43.02	28.56	5.67	9.64	6.06	14.68	25.66	15.95	54.64
	GPT	28.23	44.82	29.69	6.58	10.47	6.83	16.76	27.08	17.02	56.19
	Snap2Rev-MR	55.14	41.50	43.94	32.16	27.16	28.23	44.03	34.88	36.34	63.88
	Snap2Rev-DPO	**55.86**	41.28	43.72	**32.48**	26.93	27.99	44.26	34.47	36.11	64.14
	Snap2Rev	55.28	**42.59**	**44.34**	32.44	**27.74**	**28.54**	**44.35**	**35.49**	**36.64**	**66.12**
Arts	Qwen	16.61	31.23	24.81	3.42	3.74	2.97	14.49	18.12	13.98	50.33
	Gemma	25.90	36.19	24.40	4.05	5.89	3.93	13.27	20.89	12.65	48.76
	InternVL	26.68	41.17	26.06	5.74	9.45	5.64	14.78	24.56	15.81	52.35
	GPT	27.75	43.22	26.98	6.64	10.48	6.41	15.95	25.79	16.66	54.46
	Snap2Rev-MR	54.89	38.83	41.52	32.69	25.96	27.48	45.01	33.11	35.19	60.43
	Snap2Rev-DPO	**55.17**	38.15	41.33	**32.76**	25.83	27.38	45.11	32.61	35.08	61.23
	Snap2Rev	54.93	**39.06**	**41.85**	32.71	**26.58**	**27.84**	**45.27**	**33.53**	**35.50**	**62.62**

Table 1: Overall performance of different methods. The improvements of our method over all baselines are significant with p-value< 0.05.

where β controls the strength of the KL divergence constraint. This formulation explicitly encourages the model to assign higher probability mass to the authentic review y_p compared to the average log-probability of the negative set.

To ensure generation stability and prevent catastrophic forgetting of linguistic fluency, we integrate a regularization term based on the standard SFT loss. The final hybrid objective is defined as:

$$\mathcal{L}_{all} = \mathcal{L}_{dpo} + \lambda L_{sft}, \tag{8}$$

where λ balances preference alignment with language modeling fundamentals. This joint optimization ensures that Snap2Review not only aligns with complex user preferences but also maintains high coherence and visual fidelity.

4 Experiment

4.1 Experimental Settings

Dataset

We evaluate our proposed framework using two publicly available, real-world datasets from the Amazon review collection [1]: *Toys and Games*, and *Arts* dataset. Each data instance consists of a user ID, product ID, review text, and review images. To establish a chronological context for personalization, a user's/product's historical reviews are defined as all reviews with timestamps preceding the timestamp of the review currently being generated. Furthermore, we standardize the review length to $L = 200$ tokens, discarding reviews containing fewer than 30 words to mitigate the impact of overly short or uninformative reviews. The resulting datasets are then randomly partitioned into training (80%), validation (10%), and test (10%) sets.

Baselines

To rigorously evaluate the effectiveness of our proposed Snap2Review framework, we compare it against several strong baseline methods. First, we select representative strong Multimodal LLMs to assess the performance of general-purpose visual understanding and generation in this specific domain. The baselines include Qwen2.5-VL-3B, Gemma-3-4B, and InternVL3_5-4B, which are leading open-source models, as well as GPT-4o, representing the current state-of-the-art in commercial systems. Second, to verify the contribution of our specific modules, we formulate two ablated variants of Snap2Review. First, we evaluate w/o Retriever FT, where the domain-specific retriever is replaced with a standard off-the-shelf CLIP model. Second, we test w/o Tri-DPO (SFT only), a variant that removes the final preference alignment stage and relies solely on Supervised Fine-Tuning.

Hyper-parameters Settings

During the retrieval stage, we select the top $N = 5$ most relevant entries from the user-specific database. For the generation backbone, we employ Qwen2.5-VL-3B, a lightweight LLM chosen to ensure computational efficiency and facilitate practical deployment in real-world e-commerce scenarios. For fine-tuning, we use the AdamW optimizer with a learning rate of $2e-5$, a batch size of 2, and train for a maximum of 2 epochs, selecting the best checkpoint based on performance on the validation set. For BERTScore calculation, we use the "best-base-uncased" model. The Student t-test is applied in our experiments to test the statistical significance.

Evaluation Metrics

We employ two complementary sets of metrics to assess generation quality comprehensively. First, to evaluate lexical overlap and fluency, we report the F_1 scores of ROUGE-1, ROUGE-2, and ROUGE-L [Lin, 2004]. Following established protocols in review generation [Tang *et al.*, 2016; Li *et al.*, 2019], these metrics measure the n-gram recall between generated content and ground-truth references. Second, to assess semantic fidelity beyond surface-level phrasing, we utilize BERTScore. By computing token-level cosine similarity, BERTScore captures the deep semantic alignment between the candidate and reference reviews, ensuring a robust evaluation even when wording differs.

4.2 Overall Performance Comparison

This section presents a comparative evaluation of our model against baselines on two datasets. The results are in Table 1,

[1] https://amazon-reviews-2023.github.io/

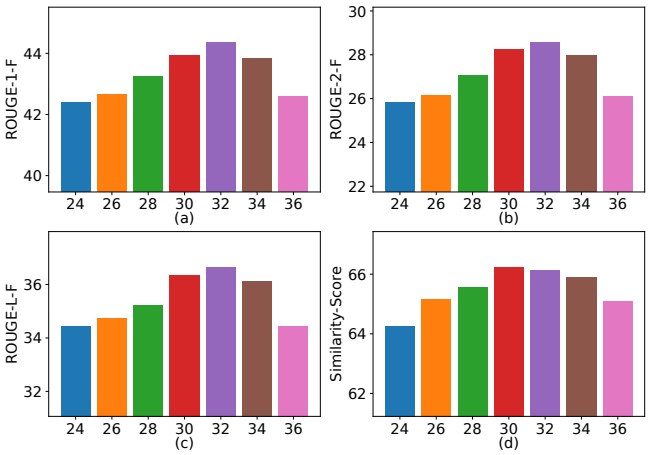


Figure 2: Performance under training steps scaled by 100.

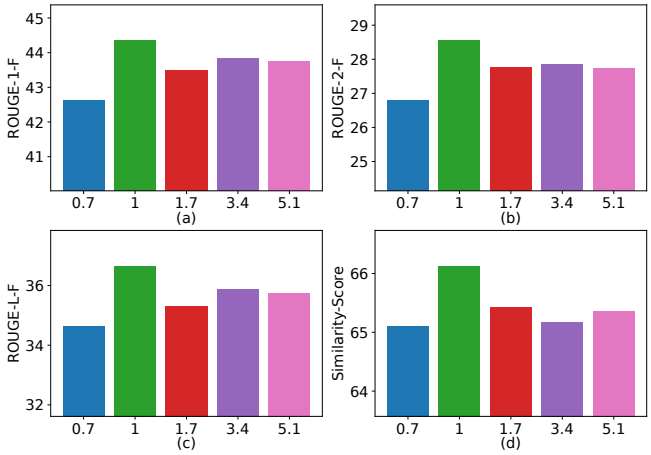


Figure 3: Performance under different loss weights.

where P, R, and F1 denote Precision, Recall, and F1-score, respectively. And we could have the following observations.

Snap2Review consistently outperforms all general-purpose MLLM baselines (e.g., Qwen, InternVL, GPT) by a substantial margin. Notably, while strong baselines like GPT achieve relatively high BertScores, their ROUGE scores remain significantly lower compared to our method. This discrepancy reveals a critical insight: general MLLMs can capture the broad semantic meaning of the product (reflected in decent BertScores), but they struggle to mimic the user's specific lexical patterns and phrasing (reflected in low ROUGE). They tend to generate generic, "AI-like" descriptions that are semantically correct but stylistically impersonal. In contrast, Snap2Review achieves high scores in both metrics, indicating that it not only understands the visual content but also accurately replicates the authentic voice of the user.

Comparing the full Snap2Review model with its variants reveals the distinct role of each module in the generation pipeline. The variant Snap2Rev-MR (w/o retriever fine-tuning) exhibits the lowest BertScore among our variants, indicating that without domain-specific alignment, the retriever fetches less relevant evidence, directly degrading the semantic fidelity of the generated content. Conversely, Snap2Rev-DPO (w/o Tri-DPO) suffers a more pronounced drop in ROUGE compared to its BertScore. This trend offers a critical insight: while the SFT baseline (Snap2Rev-DPO) can maintain general semantic relevance (decent BertScore), it fails to mimic the user's specific lexical choices and phrasing habits(reflected in lower ROUGE). The full framework achieves the best overall performance, validating the synergy between robust visual grounding and fine-grained stylistic alignment

4.3 Performance of CLIP finetuning

To evaluate the effectiveness of our retrieval optimization strategies, we compare the performance of the original Multimodal Retriever (MR) against variants with Full Fine-tuning (MR+FT) and our proposed Asymmetric Parameter Update (MR+AFT). The results are shown in Table 2.

Specifically, the off-the-shelf MR yields suboptimal results

Methods	Rec@1↑	Rec@5↑	Rec@10↑	MeanRank↓
MR	15.00	45.00	61.00	14.22
MR+FT	18.00	51.50	65.50	11.45
MR+AFT	19.50	57.50	71.00	8.72

Table 2: Performance of multimodal retriever (MR) finetuning.

due to the domain shift between general pre-training and e-commerce data. Comparing the fine-tuning strategies, the superior performance of MR+AFT over MR+FT reveals that user-provided images often contain significant visual noise (e.g., clutter, poor lighting) that can distort the embedding space during full fine-tuning. By freezing the text encoder as a robust semantic anchor, our asymmetric strategy forces the vision encoder to selectively extract meaningful features consistent with textual descriptors, effectively filtering out visual ambiguities. This precise visual-textual alignment is pivotal, as it prevents the retrieval of irrelevant evidence caused by visual ambiguity, thereby safeguarding the generator against hallucinations and ensuring the final reviews are strictly grounded in visual reality.

4.4 Influence of hyperparameters

Impact of Tri-DPO Training Steps

In this section, we verify the sensitivity of our model to the number of training steps during the Tri-DPO optimization phase. Figure 2 illustrates the performance trends across ROUGE and Similarity metrics under varying step counts.

As observed, before reaching 3200 steps, the consistent gain indicates that the model progressively masters the hierarchical curriculum, first securing basic relevance against general negative responses and subsequently resolving the finer stylistic boundaries presented by near-negative responses. However, the decline observed beyond this point points to the phenomenon of reward over-optimization. Excessive gradient updates cause the policy to drift significantly from the reference model, leading to high KL divergence where the model exploits statistical irregularities rather than learning generalizable patterns. This results in a loss of linguistic fluency, effectively degrading test performance. Consequently,

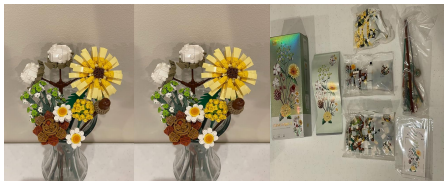
Images provided by users	Generated reviews	
	**Qwen+RAG:** I must say, I am thoroughly impressed! The flowers are incredibly detailed, and they truly look more like real flowers than plastic..... Overall, this set is perfect for anyone who enjoys building with mini bricks. It's a great Easter gift idea for brick-building enthusiasts, and it's also suitable for children who are just starting to explore building with small components. I highly recommend giving this set a try if you're looking for a fun and engaging building activity.	**Our method:** These are nice little flowers with great detail. I like the look of these more than the name brand variety, because of the finer look, they just look more like flowers than plastic. The instructions are easy to follow and I didn't have any missing pieces. The pieces fit tightly together, but the bushy yellow flower would have pieces pop out if you tried to stack them on top of each other.
	**Qwen+RAG:** I recently purchased the August doll from Babeside... The doll is incredibly realistic, with detailed features like mottling, veining, and blushing that give her a lifelike appearance..... The doll's head and limbs are made from soft vinyl, which makes her feel comfortable to hold and play with. The body is cloth, adding to the authenticity of the experience. ..... The doll's hair is painted on very nicely, and the overall design is both charming and realistic.	**Our method:** This baby boy doll is amazing for the price. He looks so real! His hair is mohair and is very soft. His outfit is so adorable too. He is a good size and his cloth body is weighted. His limbs and head are soft vinyl, so he is perfect to snuggle. He comes with a lion lovey, a magnetic pacifier, bottle, and a mat. He is a great gift for any child or collector.

Figure 4: Cases of generated reviews of our method and LLMs.

we identify 3200 steps as the optimal equilibrium, effectively balancing aggressive stylistic alignment with the preservation of natural language capabilities.

Impact of Loss Balancing Weight

We further investigate the impact of the balancing weight λ, which regulates the critical trade-off between stylistic preference alignment and foundational language capabilities. Figure 3 illustrates the performance trends across various weight configurations.

As observed, the model achieves optimal performance when the two objectives are balanced equally ($\lambda = 1$). When the weight is skewed towards preference alignment (lower λ), the reduced regularization fails to constrain the model within a plausible language manifold, causing generation instability. Conversely, when the weight heavily favors supervised learning (higher λ), we observe a notable performance degradation. This suggests that an overly strong constraint on generic token likelihood suppresses the subtle personalization signals essential for mimicking user personas. Therefore, an equal weighting provides the ideal stability, allowing the model to internalize distinct stylistic traits without compromising general fluency.

4.5 Case studies

To intuitively verify the contribution of our alignment strategy, we present a qualitative comparison in Figure 4 between the full Snap2Rev model and its ablation variant Qwen+RAG.

In the first case (top row), Qwen+RAG adopts a generic, promotional tone (highlighted in blue), utilizing phrases like "thoroughly impressed" and "highly recommend" that resemble a marketing script. In contrast, Snap2Rev captures the authentic voice of a user (highlighted in red), making specific comparisons ("more than the name brand variety") and offering personal aesthetic judgments ("just look more like flowers"), reflecting a genuine consumer perspective. The second case (bottom row) highlights a critical difference in fine-grained visual grounding. Qwen+RAG suffers from visual hallucination, incorrectly stating that "The doll's hair is painted on." Conversely, Snap2Rev accurately identifies the specific material, describing the hair as "mohair and very soft." These examples demonstrate that while RAG ensures factual relevance, Tri-DPO is the key driver that steers the model away from generic descriptions towards truly personalized, human-like expression.

5 Conclusion

In this paper, we present Snap2Review, a visually driven framework designed to bridge the gap between objective visual signals and subjective user personas in automated review generation. To tackle the persistent challenges of visual hallucination and stylistic homogenization, we decomposed the task into two synergistic stages. First, we established a visual-anchored retrieval mechanism, optimizing a cross-modal retriever to map user images to precise textual evidence, thereby ensuring a strict grounding of the generation in visual reality. Second, to capture the nuances of individual writing styles, we introduced Tri-DPO, a novel alignment algorithm leveraging a stratified negative sampling strategy. By learning to distinguish between obvious errors, stylistic mismatches, and near-authentic negatives, our model effectively moves beyond generic praise to mimic authentic user voices. Extensive experiments demonstrate that Snap2Review sets a new state-of-the-art, generating reviews that are both visually accurate and highly personalized. In future work, we plan to extend this paradigm to interactive dialogue systems and explore its generalization across diverse e-commerce categories.

Acknowledgements

This work was supported by the Outstanding Youth Team Project of Central Universities (No. QNTD202504), the National Natural Science Foundation of China (Grant Nos. 62502036, 62572058, and U24A20244), the China Postdoctoral Science Foundation (Grant No. 2024M750199), and the Fundamental Research Funds for the Central Universities (No. PTJS202635, No. BFUKF202616).

References

[Dong *et al.*, 2017] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. Learning to generate product reviews from attributes. In *ECAL*, pages 623–632, 2017.

[Dong *et al.*, 2025] Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 4206–4225, 2025.

[Fan *et al.*, 2024] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, 2024.

[Han *et al.*, 2025] Dongge Han, Trevor McInroe, Adam Jelley, Stefano V Albrecht, Peter Bell, and Amos Storkey. Llm-personalize: Aligning llm planners with human preferences via reinforced self-training for housekeeping robots. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1465–1474, 2025.

[Huang *et al.*, 2024] Qiushi Huang, Xubo Liu, Tom Ko, Bo Wu, Wenwu Wang, Yu Zhang, and Lilian Tang. Selective prompting tuning for personalized conversations with llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16212–16226, 2024.

[Huang *et al.*, 2026] Hua Huang, Qiyao Peng, Minglai Shao, and Hongtao Liu. Insightrec: Enhancing sequential recommendation through reasoning-aware preference optimization. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 20451–20455. IEEE, 2026.

[Kim *et al.*, 2020] Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, 2020.

[Li *et al.*, 2019] Junyi Li, Wayne Xin Zhao, Ji-Rong Wen, and Yang Song. Generating long and informative reviews with aspect-aware coarse-to-fine decoding. In *ACL*, pages 1969–1979, 2019.

[Li *et al.*, 2020] Junyi Li, Siqing Li, Wayne Xin Zhao, Gaole He, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. Knowledge-enhanced personalized review generation with capsule graph neural network. In *CIKM*, pages 735–744, 2020.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[Ni and McAuley, 2018] Jianmo Ni and Julian McAuley. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *ACL*, pages 706–711, 2018.

[Peng *et al.*, 2025] Qiyao Peng, Hongtao Liu, Hua Huang, Jian Yang, Qing Yang, and Minglai Shao. A survey on LLM-powered agents for recommender systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, November 2025. Association for Computational Linguistics.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[Shankar *et al.*, 2024] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14, 2024.

[Tang *et al.*, 2016] Jian Tang, Yifan Yang, Sam Carton, Ming Zhang, and Qiaozhu Mei. Context-aware natural language generation with recurrent neural networks. *arXiv preprint arXiv:1611.09900*, 2016.

[Trivedi *et al.*, 2025] Prashant Trivedi, Souradip Chakraborty, Avinash Reddy, Vaneet Aggarwal, Amrit Singh Bedi, and George K Atia. Align-pro: A principled approach to prompt optimization for llm alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27653–27661, 2025.

[Truong and Lauw, 2019] Quoc-Tuan Truong and Hady Lauw. Multimodal review generation for recommender systems. In *The World Wide Web Conference*, pages 1864–1874, 2019.

[Yang *et al.*, 2024] Dayu Yang, Fumian Chen, and Hui Fang. Behavior alignment: A new perspective of evaluating llm-based conversational recommendation systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2286–2290, 2024.

[Yu *et al.*, 2025] Tao Yu, Yi-Fan Zhang, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, et al. Aligning multimodal llm with human preference: A survey. *arXiv preprint arXiv:2503.14504*, 2025.

[Zang and Wan, 2017] Hongyu Zang and Xiaojun Wan. Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 168–177, 2017.

[Zhang *et al.*, 2025] Haobo Zhang, Qiannan Zhu, and Zhicheng Dou. Enhancing reranking for recommendation with llms through user preference retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 658–671, 2025.