Computer Science
&
Data Science

Capstone Report - Fall 2025

Collaborative Learning with Action-aware Image-text Representation Optimization

Zhaodong Liu,
Yuquan Hu,
Tuoye Liu

supervised by
Hongyi Wen

Preface

This report presents the research conducted for the Computer Science & Data Science Senior Capstone Project at New York University Shanghai. The work addresses fundamental limitations in traditional recommendation systems through the lens of generative modeling and multimodal learning.

Traditional recommendation system frameworks, which typically rely on candidate filtering and ranking pipelines, exhibit well-documented challenges in surfacing long-tail items—the vast majority of content that receives limited exposure. This limitation stems from the inherent bias of ranking-based approaches toward popular items. The paradigm shift to Generative Recommendation offers a promising alternative by reformulating user interaction as an autoregressive generation task, enabling more equitable exploration of the item space.

The work presented in this report extends the context-aware tokenization approach introduced by ActionPiece. Specifically, this research expands the framework to incorporate multimodal capabilities by integrating visual features alongside textual representations, while maintaining computational efficiency. The investigation demonstrates that robust, generalizable performance can be achieved across diverse domains without incurring prohibitive computational overhead.

This report is intended for researchers and practitioners in the fields of recommender systems, natural language processing, and multimodal learning. The proposed approach contributes to the growing body of work on bridging Large Language Models and Recommendation Systems, with particular emphasis on practical applicability. The findings provide insights into designing efficient multimodal generative recommendation systems capable of handling complex user queries and improving item coverage.

Acknowledgements

Special thanks to Professor Wen, who provided invaluable guidance, insightful feedback, and continuous support throughout the development of this capstone project.

Special thanks to Claire Cottrill, who inspired the name of our architecture design and enriched our lives through her music.

Abstract

Traditional recommendation systems often struggle with the long-tail problem and cold-start user scenarios, failing to surface rarely seen items effectively. While Generative Recommendation offers a promising paradigm shift by treating interaction as an autoregressive generation task, existing frameworks typically lack the capability to integrate multimodal data efficiently or rely on independent tokenization that ignores contextual relationships.

To address these limitations, we propose CLAIRO (Collaborative Learning with Action-aware Image-text Representation Optimization). Building upon the context-aware tokenization, our approach introduces a lightweight multimodal fusion pipeline to integrate visual features efficiently. We utilize a modified token merging mechanism to jointly cluster these features, enabling the model to discover latent co-occurring patterns across modalities without significant computational overhead.

We evaluate on four categories from Amazon Review Data (2018), demonstrating that CLAIRO significantly outperforms both text-only baselines and existing multimodal models across key metrics. These results confirm that our action-aware fusion of visual and textual signals provides a more robust and equitable representation of items, successfully enhancing recommendation accuracy for vague user queries.

Keywords

Natural Language Processing; Generative Retrieval; Multimodal Learning

Contents

1 Introduction

We study how to build more effective recommendation systems using NLP techniques, with a focus on overcoming two long-standing challenges: long-tail items and cold-start users. Traditional dense retrieval recommendation system typically follow a candidate-generation–plus-ranking pipeline, which tends to over-favor popular items and ignore long-tail items [1].They also struggle with the cold-start problems, where user interaction histories are limited, leading to inconclusive, unreliable candidate filtering and poor personalization in the early stages.

To address these limitations, we propose a **Collaborative Learning system with Action-aware Image-text Representation Optimization** (CLAIRO) based on autoregressive sequence modeling. Building on the foundational context-aware tokenization of ActionPiece[2], we aim to extend multimodal capability for the model. Specifically, we use novel methods to combine visual and textual information, improving its understanding on multiple kinds of item features, while also maintaining its efficiency by applying feature integration methods to leverage the increasing computational overhead caused by excessive data of different modalities. Our goal is to make advanced generative recommendation techniques both practical and scalable for real-world industry applications, aiming for reduced time and high recommendation accuracy.

2 Related Work

These works provides a comprehensive overview of the generative recommendation task and propose several approaches to enhance the performance.

Large Language Models for Generative Recommendation[3] provides a comprehensive survey of generative approaches to recommendation systems. Traditional recommendation systems rely on multi-stage pipelines with separate candidate retrieval and ranking stages, while Generative Recommendation shows a paradigm shift by representing user and item identifiers as token sequences rather than fixed embeddings, enabling direct generation of recommendations through autoregressive models, significantly improving its efficiency. GRU4REC[4] was the first to implement GRU-based RNN for these tasks, and inspired by previous works, Transformers4Rec[5] used transformer models with masking for sequential recommendation tasks.

To elaborate on the transition from RNNs to the advanced transformer models mentioned above, it is worth noting several key milestones. A significant architectural leap was made by SASRec[6], which first successfully applied a self-attention mechanism, similar to a Transformer

decoder, to capture user behavior sequences. Building on this, and heavily inspired by the success of masked language modeling in NLP, BERT4Rec[7] further advanced the field by popularizing the use of masking strategies in training, a paradigm that directly influenced models like Transformers4Rec.

TIGER[8] introduced a method called Transformer Index for GEnerative Recommenders, which proposes a generative retrieval paradigm that represents each item with Semantic ID[9], a semantically meaningful identifier. Specifically, TIGER introduces a novel method of generating Semantic ID, which are discrete token tuples generated by a pre-trained **SentenceT5**[10] text encoder, quantizing item content embeddings using a **RQ-VAE**[11] quantizer, enabling hierarchical representation of items. Empowered by the P5[12] fine-tuning method, the Semantic ID representation of items are learned based on the content information of the items.

However, RQ-VAE[11] still has its problem. The tokenization on each action is independent, assigning the same fixed tokens to identical actions across all sequences without considering contextual relationships. **ActionPiece**[2] proposes a novel tokenization method by merging original feature patterns as new tokens. This method indeed considers the contextual relationship during the tokenization process. In detail, each action as a token consists of a set of item features, and the frequently co-occurring tokens will merge as a new token using a bottom-up method similar to Byte-Pair Encoding(**BPE**) [13]. The set permutation regularization method is also implemented to improve robustness and performance. This shows that the tokenization method of merging considering co-occurrence frequency provides a solid foundation for generative models to capture complex user–item interaction patterns, there by improving recommendation quality and efficiency.

Though the ActionPiece[2] has reached SOTA performance in the field of generative recommendation systems, it still lacks the ability to perform on multiple modal tasks. Meanwhile, for the RQ-VAE[11] based models, recent approaches have explored the integration of multiple modalities, such as text and images, to improve recommendation quality. MMGRec[14] devise a hierarchical quantization method Graph RQ-VAE to assign Rec-ID consisting of a tuple of semantically meaningful tokens for each item from its multimodal information. Then train a Transformer-based recommender to generate the Rec-IDs of user-preferred items based on historical interaction sequences. **MQL4GRec** [15] introduces a novel framework that leverages quantitative language[15] to represent multimodal item content. By utilizing RQ-VAE[11] with a translator(LLaMA[16] for text and ViT[17] for image), the system translates both textual and

visual features into discrete token sequences. These tokenized sequences serve as a unified representation of the item content, which enables the model to seamlessly process multimodal data and enhance both the diversity and personalization of recommendations. Unlike traditional methods that treat different modalities separately, MQL4GRec[15] integrates them into a single language space, making the recommendation process more efficient and scalable.

3 Solution

3.1 ActionPiece Tokenization Framework

ActionPiece[2] has introduced an advanced merge algorithm, which is a context-aware adaptation of Byte Pair Encoding (BPE)[13] designed to tokenize sequences of unordered feature sets. Starting with a vocabulary of atomic features, it iteratively merges the most frequent token pair into a new token. Crucially, it calculates merge scores using a weighted frequency that distinguishes between intra-set pairs (features within the same action) and inter-set pairs (features across adjacent actions), allowing it to capture both item attributes and sequential context. To implement this efficiently, the algorithm utilizes a double-ended linked list structure with "intermediate nodes" to manage connections and store tokens that span across boundaries, dynamically updating the corpus and vocabulary until a target size is reached.

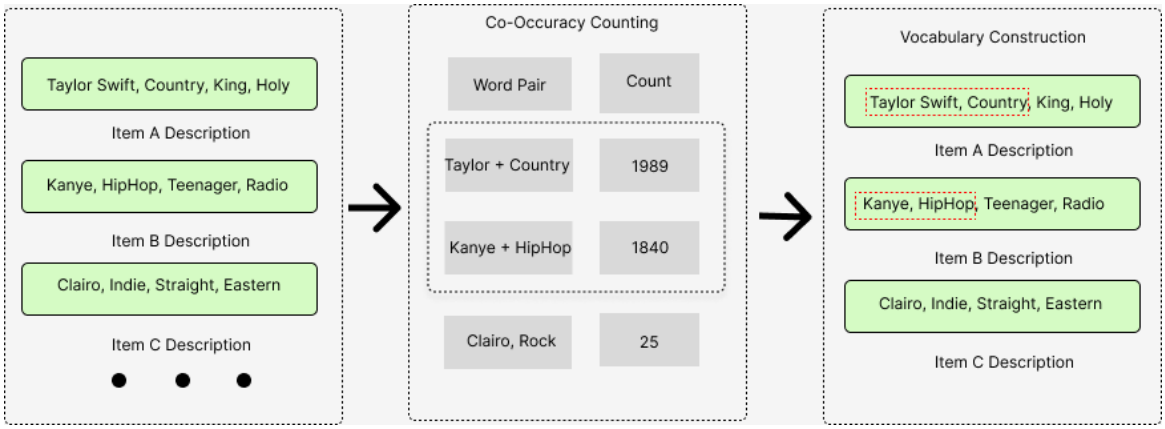

Figure 1: Illustration of Merging Algorithm of ActionPiece

3.2 Extending ActionPiece to Multimodal Tokenization

Building upon ActionPiece's[2] tokenization framework, we have successfully extended its capability to incorporate visual features. Our current implementation integrates a visual encoder to extract image embeddings for product images, which are then projected into a shared semantic space with textual features. This design allows CLAIRO to flexibly leverage complementary information from different modalities and better model complex item characteristics reflected in both appearance and description.

The key technical advancement in our implementation is the multimodal token merging module. The resulting behavioral tokens encode both textual semantics and visual appearance, providing richer item representation embedding for the recommendation task. We incorporated a CLIP ViT-L/14[18] as the visual encoder. Combining textual data with visual data, CLAIRO construct the action piece vocabulary in multimodality, and as expected, shows progress in its performance.

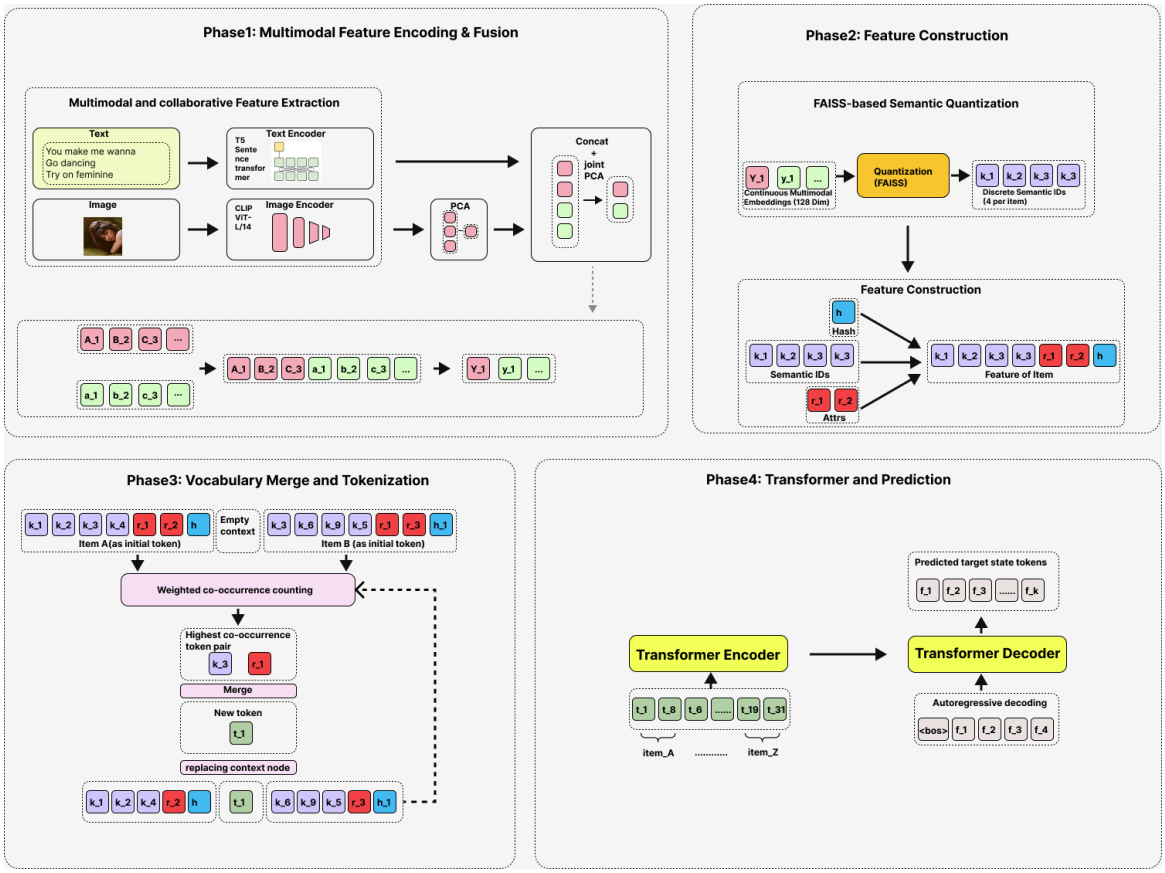

Figure 2: Pipeline of CLAIRO

3.3 Multimodal Fusion and Quantization Pipeline

To achieve this, we designed a lightweight multimodal fusion pipeline that preserves the overall ActionPiece[2] workflow while augmenting its feature extraction stage. Visual embeddings are first obtained from a pretrained ViT[17] model and then compressed via Principal Component Analysis (PCA) to unify the dimensionality of visual and textual embeddings at a 1:1 ratio. These reduced visual vectors are subsequently concatenated with the textual sentence embeddings produced by a SentenceTransformer encoder, resulting in a 768-dimensional fused vector. An optional second-stage PCA is applied to the fused representation, ensuring the multimodal embedding remains compact and well-conditioned for the subsequent quantization stages. This fused embedding is then passed into the FAISS[19]-based Optimized Product Quantization (OPQ) [20] module, where the fused 128-dimensional multimodal embeddings are decomposed into multiple subspaces. Each item is represented by a fixed-length sequence of discrete semantic codes, who will be used later in the token merging process.

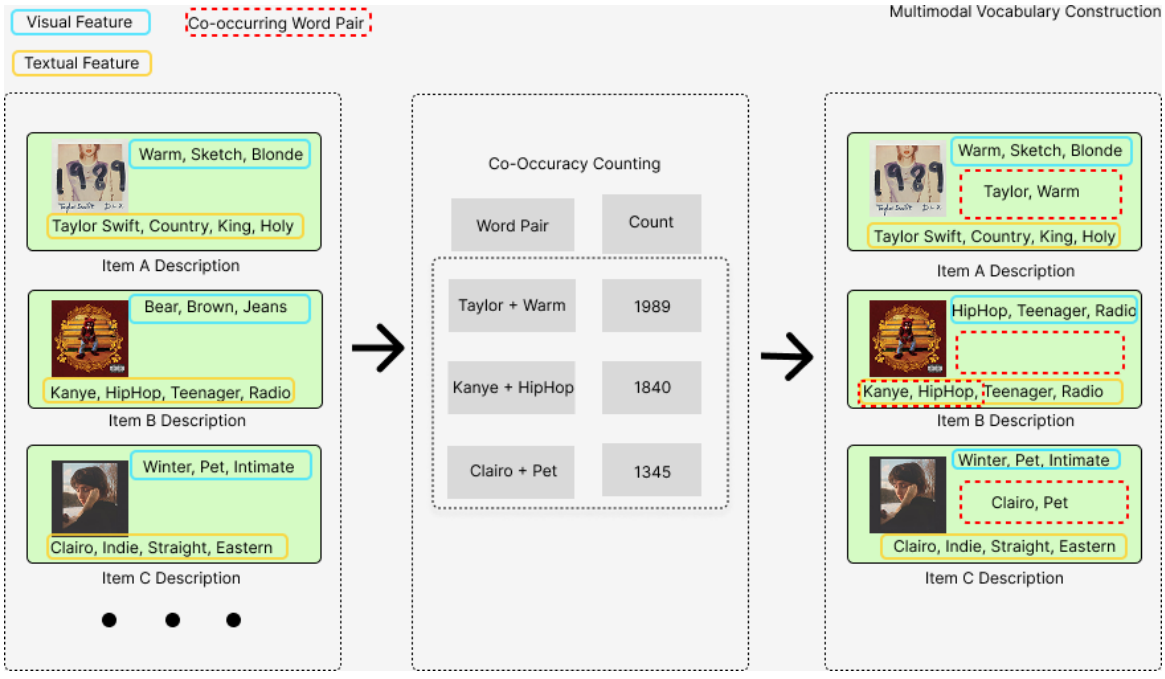

Figure 3: Illustration of Multimodal Merging Algorithm of CLAIRO

3.4 Multimodal Token Merging and Visual–Textual Co-occurrence Modeling

A central innovation of our approach lies in extending the token merging mechanism to incorporate visual tokens alongside textual tokens during vocabulary construction. Unlike prior methods that treat visual features as independently fused representations, CLAIRO integrates visual and textual signals within a unified token merging process. By jointly modeling co-occurrences between visual and textual tokens, the model is able to capture cross-modal semantic regularities that emerge when certain visual patterns consistently align with specific textual descriptions. This visual–textual co-occurring modeling enables the construction of higher-order multimodal tokens that encode collaborative semantics across modalities, rather than relying solely on late fusion or static concatenation. As a result, CLAIRO can learn more expressive and compact vocabularies that better reflect real-world item multimodal characteristics, leading to improved representation quality and downstream recommendation performance.

3.5 Dataset Compatibility and Multimodal Data Support

In addition, to provide the model with richer data in different categories, and work on more valid, reliable, and up-to-date item image feature, we also constructed a pipeline that is compatible with the newer Amazon Review Data(2018)[21] dataset, which provides more visual and textual data, enabling later multimodal adaptation.

Moreover, the newer version provides longer and denser user–item interaction histories, which allow the model to better capture sequential dependency patterns and temporal dynamics in user behavior. As a result, the enriched data foundation further strengthens the learning of collaborative representations and improves the robustness and generalization capability of the proposed approach.

4 Experimental Results

4.1 Experimentation Detail

Baselines. We have successfully replicated the baseline model and have conducted extensive experiments on them, where the performances are aligned with the performance in the paper. We also conducted additional experiments on more datasets from Amazon Review Data(2018)[21]. to demonstrate the baseline performance of ActionPiece[2] and MQL4GRec[15], allowing us to compare across the text-only baseline and multimodal textual-visual baseline with CLAIRO.

Data. We use Amazon Review Data (2018)[21] on different categories. This dataset contains 233.1 million of reviews from Amazon shopping website. Here we conduct experiments on *Sports and Outdoors*(Sports), *Musical Instruments*(Instruments), *CDs and Vinyl*(CDs), and *Arts Crafts and Sewing*(Arts) dataset categories. We believe these datasets are sufficient to evaluate the model performance and reach satisfactory results.

Evaluation. We have several different evaluation metrics to reflect on the performance of both our model and the State-of-the-Art model. Some of the methods are listed in the following.

NDCG@K (Normalized Discounted Cumulative Gain at rank K) measures not only whether the relevant items appear in the recommendation list but also how highly they are ranked within the top-K positions. A higher NDCG@K score indicates that relevant items are placed closer to the top of the list, reflecting better ranking effectiveness.

Recall@K measures the proportion of relevant items that appear within the top-K recommendations. A higher recall value shows that the model is less likely to miss items of interest to the user, which is crucial for improving user satisfaction.

Iteration time(it/s), the time of single iteration of the model reflects the efficiency of the model during inference. With a mind on efficiency and scalability, this metric could help us managing those overhead, developing a model that is efficient with low computational resource cost.

Combining those multiple metrics together provides a more comprehensive evaluation of a recommender system. NDCG@K illustrates ranking quality well, Recall@K captures coverage of relevant items. Iteration time demonstrates the time consumption of the model, indicating the efficiency. Considering all those metrics ensures a balanced assessment of both accuracy and diversity in recommendations.

4.2 Experimentation Results

Overall Performance. For CLAIRO, we have conducted multiple experiments, and validated our approach successfully. Working on Amazon Review Data[21], CLAIRO achieves improvements over the text-only baseline and multimodal textual-visual baseline on majority of Amazon Review Data categories. These improvement are statistically demonstrated across those evaluation metrics (Recall@5/10, NDCG@5/10), indicating that visual features provide complementary signals for item representation, while the action-aware token merge mechanism also works perfectly

when it encounters collaborative tokens from both visual and textual modality, demonstrating the overall robustness and effectiveness of our approach.

Upon comparison between ActionPiece[2] and CLAIRO, we found that although CLAIRO consistently outperforms ActionPiece[2] across most categories, the performance gains are generally marginal in Arts Crafts and Sewing, and Musical Instruments as shown in Table 1. The relative improvement is only around 2% - 2.5%, demonstrating a slight improvement after the integration of visual features. However, we observe significant improvements in CDs and Vinyl, and Sports and Outdoors categories, with a improvement of 30% - 40%. We speculate that the difference of visual feature across categories lead to distinct feature embedding, hence influence the performance with visual feature integration.

This observation suggests that in some categories, visual features very likely play a complementary role in the collaborative embedding. Even though visual representations constitute roughly half of the embedding dimensions, they do not fundamentally alter recommendation outcomes. Instead, they provide auxiliary signals that slightly enhance ranking quality on top of already strong textual representations.

But in other categories, the visual features play a much more important role instead, where the integration of visual feature embedding could provide useful representation to help the model better understand item feature, hence make prediction of user interaction with higher accuracy.

Dataset	**Metrics**	**ActionPiece**	**CLAIRO**	**Improv.**
Arts	Recall@5	0.1351	0.1373	1.6%
	Recall@10	0.1562	0.1597	2.2%
	NDCG@5	0.1190	0.1216	2.2%
	NDCG@10	0.1257	0.1287	2.4%
CDs	Recall@5	0.0594	0.0787	32.5%
	Recall@10	0.0832	0.0988	18.8%
	NDCG@5	0.0441	0.0640	45.1%
	NDCG@10	0.0523	0.0705	34.8%
Instruments	Recall@5	0.0977	0.1006	3.0%
	Recall@10	0.1197	0.1229	2.8%
	NDCG@5	0.0832	0.0846	1.7%
	NDCG@10	0.0902	0.0918	1.8%
Sports	Recall@5	0.0423	0.0570	34.8%
	Recall@10	0.0492	0.0635	29.1%
	NDCG@5	0.0339	0.0499	47.2%
	NDCG@10	0.0361	0.0520	44.0%

Table 1: Performance comparison between ActionPiece and CLAIRO

Compared with the multimodal baseline MQL4GRec[14], CLAIRO demonstrates significant improvements across most datasets. As shown in Table 2, CLAIRO consistently surpasses MQL4GRec[14] on Arts, CDs, and Instruments by large margins, with a significant relative improvement reaching up to 135.3% in NDCG@5 on CDs and Vinyl. These results suggest that simply incorporating visual features is not sufficient; rather, effectively aligning and merging multimodal signals is crucial. The superior performance of CLAIRO highlights the effectiveness of the proposed action-aware token merge mechanism, which enables collaborative interaction between visual and textual tokens at a finer granularity.

However, we observe that on the Sports dataset, MQL4GRec[14] performs notably better than CLAIRO across all evaluation metrics. This phenomenon is likely caused by the unique characteristics of sports-related items, where product images often contain rich contextual information (e.g., athletes, scenes, and usage environments) that is rather difficult to distinguish across items and detect co-occurrence with text descriptions. Therefore, the collaborative embedding may not align well with user action signals. This further emphasizes that model performance is highly sensitive to category-specific data distributions.

Dataset	**Metrics**	**MQL4GRec**	**CLAIRO**	**Improv.**
Arts	Recall@5	0.1028	0.1373	33.5%
	Recall@10	0.1278	0.1597	25.0%
	NDCG@5	0.0847	0.1216	43.5%
	NDCG@10	0.0927	0.1287	38.8%
CDs	Recall@5	0.0409	0.0787	92.4%
	Recall@10	0.0658	0.0988	50.2%
	NDCG@5	0.0272	0.0640	135.3%
	NDCG@10	0.0352	0.0705	100.3%
Instruments	Recall@5	0.0818	0.1006	23.0%
	Recall@10	0.0933	0.1229	31.7%
	NDCG@5	0.0763	0.0846	10.9%
	NDCG@10	0.0800	0.0918	14.8%
Sports	Recall@5	0.0724	0.0570	–
	Recall@10	0.0844	0.0635	–
	NDCG@5	0.0644	0.0499	–
	NDCG@10	0.0682	0.0520	–

Table 2: Performance comparison between MQL4GRec and CLAIRO

Overall, these experimental results demonstrate that CLAIRO effectively leverages multimodal information and user actions to improve recommendation performance. At the same time, we emphasize that due to substantial differences in dataset size, textual description quality, image

sources, and visual feature characteristics across categories, the reported results are only directly comparable within the same category of Amazon Review Data (2018)[21], rather than across different categories.

4.3 Ablation Study

Several ablation studies are conducted, in exploration to the model architecture, pipeline design, and difference between dataset categories.

Effect of Final-stage PCA before OPQ. To examine whether we should compress the collaborative embedding initially by a final-stage PCA before OPQ[20] quantization through FAISS[19] or directly send it into OPQ[20] process, we conduct experiments to compare the performance of two methods. Table 3 demonstrates the results with or without final-stage PCA and its improvement in performance. The experimental results clearly indicate that removing the final-stage PCA consistently leads to substantially better performance improvement by approximately 37%–42% on all evaluation metrics.

Metrics	**w/o Final PCA**	**w/ Final PCA**	**Improv.**
Recall@5	0.0787	0.0561	40.3%
Recall@10	0.0988	0.0713	37.2%
NDCG@5	0.0640	0.0450	42.2%
NDCG@10	0.0705	0.0499	41.3%

Table 3: CLAIRO Performance with different embedding processing on CDs and Vinyl

Contribution of Visual and Textual Embeddings. We compare the collaborative token with a text-only variant. CLAIRO demonstrates similar performance to the text-only baseline, suggesting that textual embeddings dominate the discriminative capability. However, the visual-only CLAIRO failed because too many conflicts of semantic IDs were observed, which are caused by the limited number of hash buckets.

5 Discussion and Analysis

5.1 Impact of Final-stage PCA before OPQ

This degradation caused by final-stage PCA can be explained by the overlapping and potentially conflicting roles of PCA and OPQ[20]. PCA performs global linear dimensionality reduction by

preserving directions with the largest variance, which may inadvertently discard fine-grained information that is crucial for similarity-based retrieval. In contrast, OPQ[20] explicitly learns an optimal orthogonal rotation to minimize quantization error across subspaces, already redistributing variance in a way that is tailored to product quantization. These results demonstrate that, within a FAISS-based OPQ[20] pipeline, final-stage PCA is unnecessary and even detrimental.

However, directly applying OPQ[20] without a final-stage PCA can increase computational cost. We have validated this by examining the time consumption during ActionPiece[2] vocabulary construction step, and resulted an increase. With a mind of industry employment, the increasing cost of ActionPiece[2] vocabulary construction could be amortized over future queries, and the majority time is consumed by the token merge process. Therefore, this modification of PCA would not significantly degrade the overall efficiency in application.

5.2 Limitations of Visual-only Collaborative Tokens

To examine the importance of visual and textual embedding in the collaborative token, we conduct experiments with only visual embedding, and only textual embedding. CLAIRO demonstrates similar performance to the text-only baseline as expected, since they have similar model architecture.

The problem we encountered was that too many conflicts of semantic IDs were observed. This issue is mainly caused by the limited number of hash buckets, which forces a large number of distinct semantic IDs to be mapped into the same buckets. This problem is usually solved by increasing the number of hash bucket size. However, we failed to solve that even though making bucket size 8x times as initial setting.

We speculate that this behavior stems from the limited diversity and high similarity of the visual data embeddings. As we have concluded in the comparison of text-only baseline and CLAIRO, the marginal improvement indicates that the visual features play a complementary role despite consuming half of the collaborative embedding. Since many embeddings lie close to each other in the feature space, the quantization process assigns them to very similar or identical codes, regardless of the increase in bucket size. Consequently, expanding the hash space does not significantly reduce collisions, as the underlying representations lack sufficient separability. This suggests that the collision issue is fundamentally constrained by the discriminative capacity of the visual embeddings rather than the hash configuration itself. Therefore, we conclude that the visual data embedding plays a complementary rather than dominant role in the collaborative

token, providing auxiliary semantic cues while relying primarily on other modalities to achieve effective discrimination.

5.3 Performance Difference of CLAIRO across Categories

This behavior could be attributed to the differences in visual characteristics across dataset categories. Since the embedding integrates both textual and visual modalities, its effectiveness is highly dependent on the discriminative power and consistency of visual features within each category. We speculate that in categories where item images are visually homogeneous or weakly correlated with user preferences, visual features only contribute limited additional information and thus lead to marginal performance gains. Conversely, in categories where images exhibit richer semantic diversity or capture usage-related cues not explicitly described in text, visual information can provide stronger complementary signals, resulting in more noticeable improvements.

Based on the representative product images shown in Figure 3, we can further explain the category-dependent behavior of visual features observed in our experimental results.

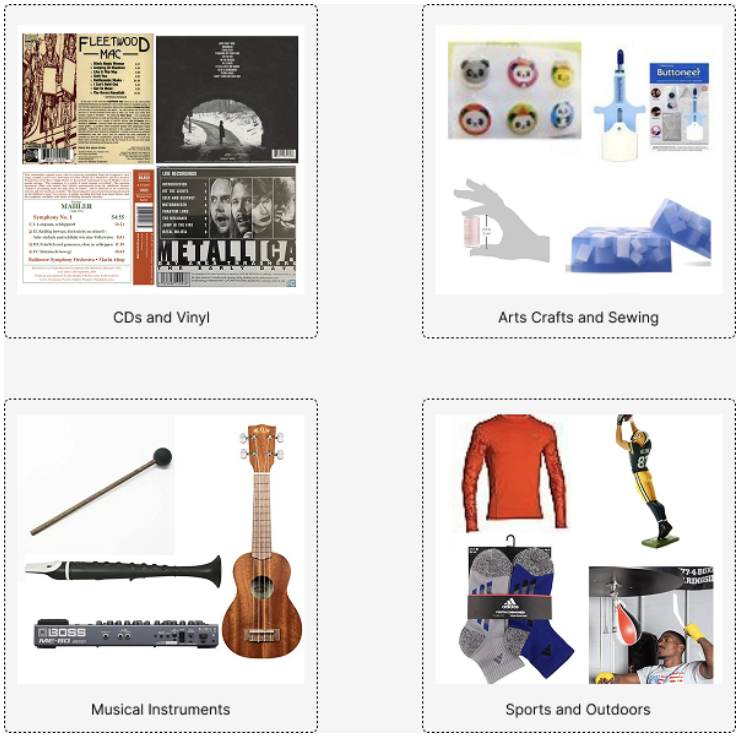

Figure 4: Selected Product image

For CDs and Vinyl, product images are predominantly album covers, which contain highly structured and semantically rich visual information, such as artistic style, genre cues, band iden-

tity, and emotional tone. These visual elements are often strongly correlated with user preferences and thus provide highly complementary signals to textual descriptions. As a result, visual features significantly enhance item representations in this category, explaining the performance gains achieved by CLAIRO. In addition, the outstanding performance gap between CLAIRO and the multimodal baseline also indicates the effectiveness of token merging algorithm with semantically rich visual feature embedding, we speculate this complexity and meaningfulness help the model to capture co-occuring visual-textual tokens.

In contrast, images in the Arts Crafts and Sewing category are considerably more diverse, ranging from raw materials and tools to functional illustrations and usage demonstrations. And for Musical Instruments, product images also tend to be highly diverse and standardized, primarily depicting different kinds of instruments against plain backgrounds. This visual heterogeneity may weaken the direct alignment between image features and user interests, making the visual modality less discriminative. Consequently, visual information contributes only marginal improvements when combined with textual embeddings.

Finally, in the Sports and Outdoors category, images include complex information, such as human actions, body poses, and environmental scenes. While such images are visually rich. With many complementary data describing the item in different aspects, the resulting features may align well with user interaction signals, introducing additional information into the embedding, thereby enhancing the co-occurring token merging process. Hence, CLAIRO surpasses the text-only baseline ActionPiece[2].

However, the rich and additional visual data that is not highly related to the textual description also bring about a problem. The token merging process depends on detecting the similarity for inter-modal and intra-modal co-occurring features, while the huge difference between visual and textual data here may defect this process. This interference would further affects the stability of the resulting textual embeddings, making it more difficult for the model to generate consistent codes and to construct a robust and informative vocabulary. Hence, CLAIRO fails to surpass the multimodal baseline.

5.4 Efficiency Analysis

With efficiency and computational overhead as primary design considerations, the architecture of CLAIRO deliberately avoids introducing excessively high-dimensional feature embeddings. Instead of scaling performance through enlarged input representations, we integrate visual features

with textual features in a compact and controlled manner, ensuring that multimodal modeling does not incur a substantial increase in computational cost. In particular, embedding compression techniques are applied during the multimodal fusion stage to limit dimensional growth while preserving the most informative signals.

Throughout our experiments, we systematically monitor the efficiency of CLAIRO using **iteration time** (it/s) as a unified metric that reflects embedding computation cost. By carefully controlling the number of discrete semantic codes generated by the FAISS-based quantization module, we prevent an explosion in vocabulary size and ensure that the complexity of the token-merging process remains stable, even as multimodal embeddings are introduced. This design choice allows CLAIRO to maintain a computational profile comparable to that of the text-only baseline while benefiting from richer item representations.

Moreover, we identify vocabulary construction as a possible computational bottleneck of the system. Since with large amount of embedding, computation in quantization process would increase a lot. To address this issue, we explore potential optimization strategies during vocabulary construction. One such attempt is the optional final-stage PCA, which reduces embedding dimensionality prior to quantization, thereby lowering the computational burden of OPQ quantizer. Although this technique offers a efficiency gains, our ablation study results indicate that it introduces a trade-off between speed and recommendation performance, highlighting the need for more principled optimization strategies in future work.

6 Future Work

6.1 Align Other Modalities

As a multimodal recommendation system, we also plan to incorporate additional data modalities, including video and audio, to further enrich item representations. These modalities naturally encode temporal dynamics, acoustic patterns, and fine-grained contextual cues that cannot be fully captured by static images or textual descriptions alone. By projecting heterogeneous inputs from different modalities into a unified token space, CLAIRO can model cross-modal interactions more effectively and capture deeper semantic correlations among modalities through the co-occurring token merging mechanism.

The integration of richer modalities is expected to be particularly beneficial for certain Amazon Review Data categories, such as CDs and Vinyl, where audio signals can directly reflect

musical style and content, and videos may convey performance context or production quality. Such modality-aware representations can complement existing visual and textual embeddings, enhancing the expressiveness and robustness of collaborative embeddings. Ultimately, incorporating additional modalities enables the recommendation system to better model complex user preferences, reduce modality-specific ambiguity, and improve recommendation performance in increasingly multimedia-driven real-world environments.

6.2 Improve Visual-only Data Performance

Looking ahead, primary focus will be on addressing the semantic redundancy observed in visual-only tokenization, where limited embedding diversity led to frequent hash collisions; we aim to mitigate this by exploring contrastive learning techniques or domain-specific fine-tuning that pushes item representations distinctively apart in the latent space. To achieve this, data cleaning will be processed to ensure the training data, especially images to match the original item information well and contain enough information to be distinguished in the quantization stage. We further conjecture that the limited discriminative capacity of visual embeddings may stem from the visual encoder itself; therefore, future work will also investigate enhanced visual encoders or task-adaptive training schemes that enable the model to generate embeddings with stronger sensitivity to fine-grained visual differences.

6.3 Explore Higher-Performance Visual and Textual Encoders

While our current implementation leverages CLIP ViT-L/14 [18] as the visual encoder and SentenceT5 [10] as the textual encoder, we plan to explore alternative high-capacity encoder architectures to further enhance the representation quality of CLAIRO. More expressive encoders are expected to capture finer-grained semantic, structural, and contextual characteristics from both visual and textual inputs, thereby producing more discriminative embeddings. Such improvements may lead to stronger alignment between modalities and more effective interaction during the co-occurring token merging process. In particular, stronger encoders could reduce modality-induced noise and improve representation stability, which is especially important for categories where visual or textual signals are weak or highly heterogeneous. This direction may further amplify the benefits of the proposed action-aware fusion mechanism and improve overall recommendation performance.

6.4 Adjust Rate of Modality in Construction

We plan to explore mechanisms that dynamically adjust the relative contribution of different modalities during embedding construction. Given the observed variability in the effectiveness of visual features across different categories, such a mechanism would allow the model to flexibly balance textual and visual information, selectively emphasizing informative embeddings while suppressing plain or noisy ones. By adaptively regulating modality weights based on category characteristics, action signals, or confidence measures, the model can better preserve discriminative textual structures while incorporating visual cues only when they provide complementary value. This adaptive modality control is expected to reduce noise propagation in the collaborative embedding space, stabilize the co-occurring token merging process, and ultimately improve the robustness and generalization ability of the learned representations across diverse recommendation scenarios.

6.5 Enable Dynamic Vocabulary Expansion

In our opinion, another promising direction for future work is to enable dynamic vocabulary expansion during training or deployment. In the current implementation, the token vocabulary is constructed offline at the very beginning and remains unchanged once the merging process is completed. While this design ensures stability and efficiency, it limits the model's ability to adapt to newly emerging interaction patterns, evolving item attributes, or unseen multimodal concepts. Introducing a mechanism for dynamically incrementing the vocabulary would allow the system to continually incorporate new high-frequency token co-occurrences without retraining the entire model from scratch. These new-coming patterns could be incrementally merged into new tokens and appended to the vocabulary, without constructing the vocabulary from the bottom ever after a change of dataset. This dynamic vocabulary expansion would enable CLAIRO to remain responsive to distribution shifts in item content and user behavior, thereby improving long-term adaptability and robustness in real-world recommendation scenarios.

6.6 Adapt to More Datasets

With our current work done entirely based on Amazon Review Data(2018)[21] dataset, we have witnessed the significant performance difference between different categories and different models. Based on the results, we speculate that the multimodal token merging process works well with semantically rich visual data. To find its logic, we aspire to expand CLAIRO into more domains

with various kind of data. Implementation on additional datasets from different domain usually contains totally different feature of Visual and Textual data.

For example, the MovieLens[22] dataset contains the movie description as textual data, and we could implement CLAIRO with movie poster or selected frame inside the movie as the visual feature source, exploring CLAIRO's ability to capture large image with complex scene and aesthetic design. We could also adapt CLAIRO to Steam Video Game and Bundle Data dataset[23], where the textual description of game, as well as various design of UI, map, and mechanics in game play as a visual data would also be challenging. We would also run experiments on yelp dataset[24], to examine the ability for the visual encoder to generate embedding for food which has diverse color and shape.

In addition to the large variety of visual and textual feature characteristics of different datasets, the different pattern of user interaction between online shopping, film watching, gaming, and dining would also be a interesting part to investigate on. People tend to purchase food in a geographically restricted area, so the dining data pattern may illustrates strong preference on local restaurants, and the taste would differ significantly for user in different areas. These additional information should all be taken into account for improving the model performance accordingly.

7 Conclusion

In this paper, we introduce CLAIRO, a novel generative recommendation framework that extends context-aware tokenization to the multimodal setting. Motivated by the fundamental limitations of traditional candidate-ranking pipelines in handling long-tail item distributions and cold-start scenarios, we reformulate the recommendation problem as an autoregressive generation task enriched with both textual and visual semantics.

Instead of directly modeling item IDs or continuous embeddings, CLAIRO represents user behaviors as sequences of discrete, structured tokens. These tokens are constructed by merging heterogeneous feature patterns derived from both textual descriptions and visual appearances, enabling the model to capture high-level semantic regularities across modalities. During vocabulary construction, we compute the weighted co-occurrence statistics of token pairs, where both intra-item feature dependencies and inter-item sequential relations are considered. The most frequently co-occurring token pairs are iteratively merged into new higher-order tokens, progressively forming a compact and expressive vocabulary.

To enable scalable and efficient vocabulary construction, we employ a double-ended linked list structure to dynamically maintain the evolving token sequences, while intermediate nodes are introduced to store merged tokens that span across feature boundaries. This design avoids repeated global re-scanning of the corpus and ensures efficient updates throughout the iterative merging process.

To further extend ActionPiece[2] to multimodal data, CLAIRO incorporates a lightweight multimodal fusion pipeline that integrates visual embeddings with textual representations into a shared semantic space. Visual embeddings extracted from a pretrained vision encoder are first compressed via PCA-based dimensionality reduction, then concatenated with sentence-level textual embeddings. An optional second-stage PCA is applied to ensure a compact and well-conditioned fused representation before quantization. Through this design, CLAIRO successfully constructs collaborative multimodal tokens that jointly encode both visual appearance and textual semantics, allowing the model to capture latent co-occurring behavioral patterns across modalities.

References

[1] L. Yang, F. Paischer, K. Hassani, J. Li, S. Shao, Z. G. Li, Y. He, X. Feng, N. Noorshams, S. Park, B. Long, R. D. Nowak, X. Gao, and H. Eghbalzadeh, "Unifying generative and dense retrieval for sequential recommendation," 2024. [Online]. Available: https://arxiv.org/abs/2411.18814

[2] Y. Hou, J. Ni, Z. He, N. Sachdeva, W.-C. Kang, E. H. Chi, J. McAuley, and D. Z. Cheng, "Actionpiece: Contextually tokenizing action sequences for generative recommendation," *arXiv preprint arXiv:2502.13581*, 2025.

[3] L. Li, Y. Zhang, D. Liu, and L. Chen, "Large language models for generative recommendation: A survey and visionary discussions," in *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[4] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.06939

[5] G. de Souza Pereira Moreira, S. Rabhi, J. M. Lee, R. Ak, and E. Oldridge, "Transformers4rec: Bridging the gap between nlp and sequential / session-based recommendation," in *Proceedings of the 15th ACM Conference on Recommender Systems*, ser. RecSys '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 143–153. [Online]. Available: https://doi.org/10.1145/3460231.3474255

[6] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," 2018. [Online]. Available: https://arxiv.org/abs/1808.09781

[7] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," 2019. [Online]. Available: https://arxiv.org/abs/1904.06690

[8] S. Rajput, N. Mehta, A. Singh, R. H. Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Q. Tran, J. Samost, M. Kula, E. H. Chi, and M. Sathiamoorthy, "Recommender systems with generative retrieval," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[9] Y. Hou, Z. He, J. McAuley, and W. X. Zhao, "Learning vector-quantized item representation for transferable sequential recommenders," 2023. [Online]. Available: https://arxiv.org/abs/2210.12316

[10] J. Ni, G. H. Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," 2021. [Online]. Available: https://arxiv.org/abs/2108.08877

[11] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," 2021. [Online]. Available: https://arxiv.org/abs/2107.03312

[12] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)," in *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*, 2022, pp. 299–310.

[13] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162/

[14] H. Liu, Y. Wei, X. Song, W. Guan, Y.-F. Li, and L. Nie, "Mmgrec: Multimodal generative recommendation with transformer model," 2024. [Online]. Available: https://arxiv.org/abs/2404.16555

[15] J. Zhai, Z.-F. Mai, C.-D. Wang, F. Yang, X. Zheng, H. Li, and Y. Tian, "Multimodal quantitative language for generative recommendation," 2025. [Online]. Available: https://arxiv.org/abs/2504.05314

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[19] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2025. [Online]. Available: https://arxiv.org/abs/2401.08281

[20] T. Ge, K. He, Q. Ke, and J. Sun, "Optimized product quantization," Tech. Rep. MSR-TR-2013-59, May 2013, iEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). [Online]. Available: https://www.microsoft.com/en-us/research/publication/optimized-product-quantization/

[21] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[22] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, Dec. 2015. [Online]. Available: https://doi.org/10.1145/2827872

[23] A. Pathak, K. Gupta, and J. McAuley, "Generating and personalizing bundle recommendations on steam," in *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2017, pp. 1073–1076.

[24] Yelp Inc., "Yelp open dataset," 2023. [Online]. Available: https://www.yelp.com/dataset