Multimodal Product Search and Retrieval Using Image–Text Representations

Project Report

Team Members:
Teera Yong
Tharuni Tekula
Swetha Narasimhan
Group 2-2

Instructor: Christopher Dunham

Date: 12/09/2025
Syracuse University School of Information Studies
IST 691 – Deep Learning in Practice

Project Overview

This project focuses on building a multimodal product search system that can understand both images and natural-language queries for an e-commerce catalog based on the Amazon Berkeley Objects (ABO) dataset (Zhai et al.). Instead of relying on basic keyword matching, we choose SigLIP ViT-B/16, a large vision-language transformer model originally trained on about 10 billion image–text pairs, to map product photos and their descriptions into a shared embedding space where semantically similar items are close together (Amazon). We then fine-tune SigLIP on a curated subset of ABO so its representations adapt to the specific visual styles and vocabulary of home and fashion products, then indexes these embeddings with a FAISS similarity search engine to support fast text-to-image and image-to-text retrieval. In a production setting, this system is intended to answer queries with low latency while remaining robust to noisy or ambiguous inputs and scalable to catalogs with millions of items, ultimately improving product discovery and laying the groundwork for personalization or recommendation features that build on the same multimodal embeddings.

Prediction, Inference, and Other Goals

The primary prediction of this project is multimodal retrieval. Given either a text query or an image, the system computes similarity scores between the query and a large set of candidate products, then ranks those products so that the correct or most relevant products appear near the top (Manning et al.). We are also interested in how fine-tuning the SigLIP model changes the shared embedding space for images and text: whether the embeddings of the same product type become more tightly clustered or scattered and whether the embeddings of different product types are better separated. This helps us understand not only how well the system retrieves, but also how it organizes multimodal information internally.

Although the current project focused on offline inference and predictions, the system is designed to support additional improvements in the future. The same system could be made faster, more robust, and more personalized which includes real-time inference, handling large-scale product catalogs, robustness to malformed or multilingual queries, personalization through user history embeddings, multi-modal reranking, and cold-start handling for new products.

Data Exploration

The Amazon Berkeley Objects (ABO) dataset contains 147,000 products across 576 categories with rich metadata including item name, brand, product type, bullet points, color, material, product descriptions, item keywords, many of which are stored as lists of language-tagged values (Amazon). In addition to textual metadata, ABO provides a separate collection of image metadata with image IDs, file paths, and image dimensions, and each product is associated with multiple product images. In terms of product category distribution, ABO dataset is heavily imbalanced; for instance, CELLULAR_PHONE_CASE accounts for 64,853 products, SHOES for 12,965 products, and GROCERY and HOME for 6,546 and 5,264 products, respectively, while many other categories have significantly fewer products.

In this project, we use only a subset of the entire ABO dataset to fine-tune our SigLIP model. We filter the dataset so that key fields such as brand, item name, bullet points, and color are available in English. This ensures that the text the model sees is coherent and meaningful for English queries. We also restrict the data to 15 product types within a home/fashion theme that are in a similar range of products, in order to help minimize the effect of class imbalance. These product types

include: HOME, CHAIR, HOME_FURNITURE_AND_DECOR, SOFA, HOME_BED_AND_BATH, RUG, FINENECKLACEBRACELETANKLET, HANDBAG, FINERING, BOOT, TABLE, FINEEARRING, WALL_ART, SANDAL, and LIGHT_FIXTURE. Because some products have multiple images, and the associated images are often less informative (e.g., close-up corners, diagrams, packaging, or accessory views), we primarily select the main image for each product. Together, these design choices result in 11,152 products, each represented by a single image.

Methods

Dataset

Before fine-tuning the model, the dataset of 11,152 products must be preprocessed. On the text side, this involves extracting string values from nested fields, cleaning them (lowercasing, removing underscores, and trimming extra whitespace), and then concatenating the 8 key fields (item name, brand, product type, bullet points, color, material, product description, and keywords) into a single, informative description. This combined caption captures both high-level attributes (such as product type, keywords) and fine-grained details (such as materials, color, and bullet points). On the image side, we join the product listings with the image metadata, construct the full file path for each product image, and remove duplicates. We also apply several image augmentations, including random resized cropping to 224×224, random horizontal flips, color jitter, small rotations, and perspective transforms, followed by normalization with SigLIP's standard mean and standard deviation. Finally, the dataset is split into 80% training, 10% validation, and 10% test sets.

Model Development and Retrieval Pipeline

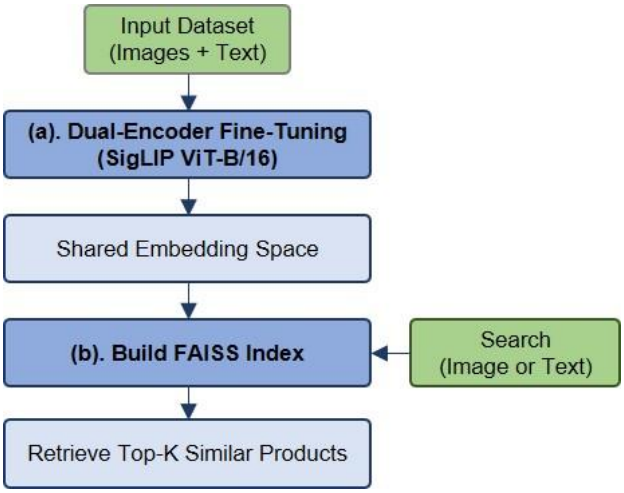

Figure 1: End-to-End Multimodal Retrieval Pipeline

The end-to-end retrieval pipeline, starting from the input dataset, through dual-encoder fine-tuning and embedding generation, to FAISS-based retrieval of top-K similar products is illustrated in **Figure 1**. First, we use the paired image-text dataset to fine-tune a dual-encoder SigLIP ViT-B/16

model, which maps both modalities into a shared embedding space where matching image–text pairs are close together. We then build a FAISS index over these embeddings. At query time, either an image or a text is encoded by the fine-tuned model and searched against the FAISS index to retrieve the top-K most similar products.

Model Fine-Tuning:

All trainable weights in both the vision and text encoders are unfrozen to perform fully fine-tuning, allowing the model to adapt the entire representation space to the ABO products and their text description. The only parameter kept frozen is the internal logit scale used to rescale cosine similarities before applying the contrastive loss; this value is already well-calibrated during pretraining and does not require further learning. We fine-tune the model over 30 epochs using AdamW optimizer with a learning rate of 1e-4 and weight decay of 0.01 to prevent overfitting while still allowing meaningful updates to the pretrained weights. During each training step, image and text embeddings are L2-normalized, a similarity matrix is computed via inner product (equivalent to cosine similarity for normalized vectors), and a symmetric contrastive loss is applied in both image-to-text and text-to-image directions (Zhai et al.; Radford et al.). This encourages matched image–text pairs to have high similarity and non-matching pairs to be pushed apart, sharpening the structure of the joint embedding space.

Retrieval with Similarity Search:

Once fine-tuning is complete, we extract embeddings for all product images and their text descriptions, creating a joint representation space where semantically related image–text pairs lie near each other. These embeddings are stored in separate similarity search indexes using Facebook AI Similarity Search (FAISS), which performs fast inner-product (cosine) similarity search (Johnson, Douze, and Jégou). At inference time, a user-provided text query is encoded by the SigLIP text encoder into a normalized embedding and searched against the image index to retrieve the top-K most similar product images; conversely, a user-provided image is encoded by the vision encoder and searched against the text index to return the most relevant product descriptions and entries. To evaluate retrieval performance, we use the test set to compute a similarity matrix and derive text-to-image Recall@1, Recall@3, and Mean Reciprocal Rank (MRR).

Results

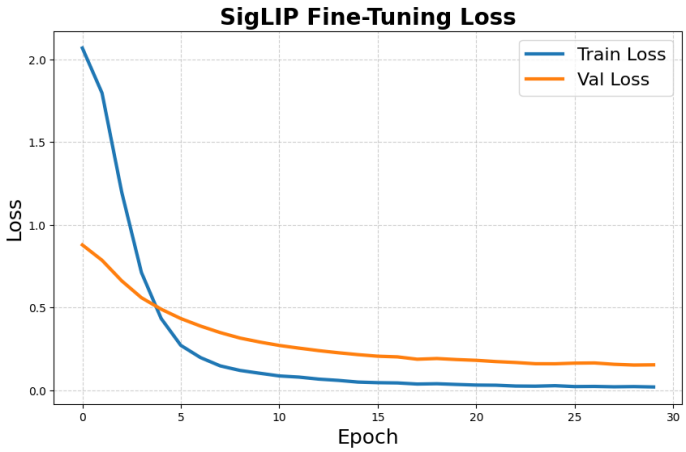

Figure 2: Training and Validation Loss Curves for SigLIP Fine-Tuning

Figure 2 illustrates the training and validation loss curves over 30 epochs of fine-tuning the SigLIP model. The training loss decreases rapidly during the first few epochs, indicating that the model quickly adapts to the ABO dataset. After this initial drop, the curve continues to decline smoothly and level off after 13 epochs. The validation loss follows a similar downward trend, though at a slightly higher magnitude, suggesting good generalization without signs of overfitting. The steady convergence of both curves demonstrates stable optimization and confirms that the chosen hyperparameters, learning rate schedule, and data augmentations support effective fine-tuning of the dual-encoder model.

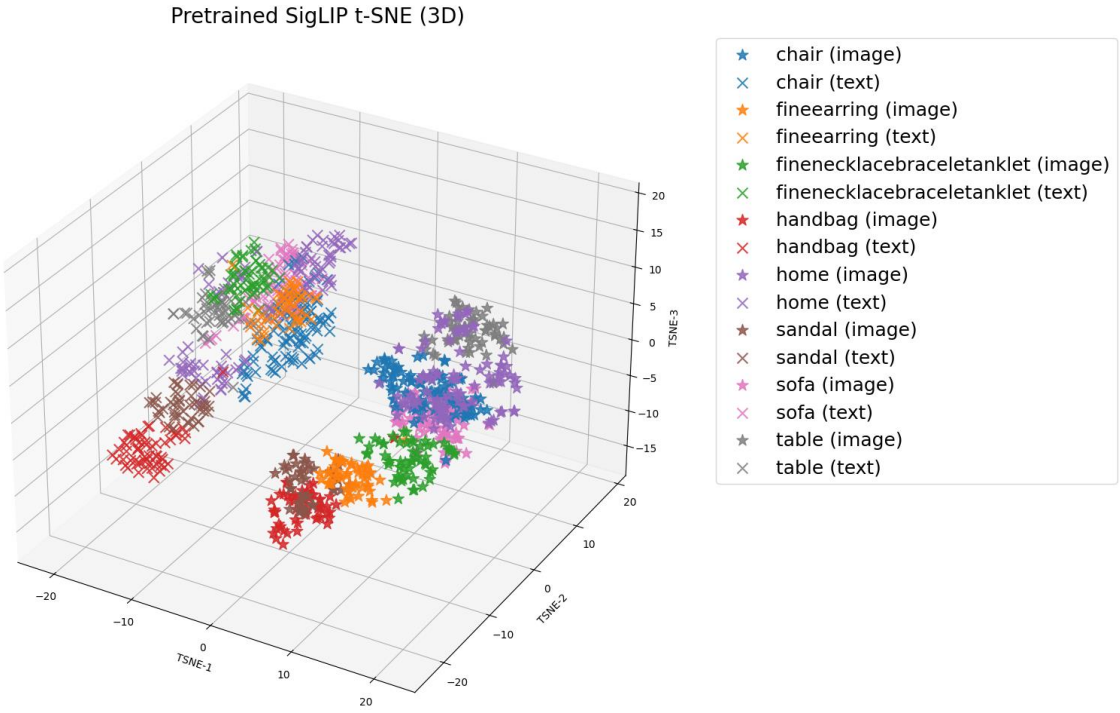

Figure 3: 3D t-SNE of Pretrained SigLIP Embeddings for Eight Product Types

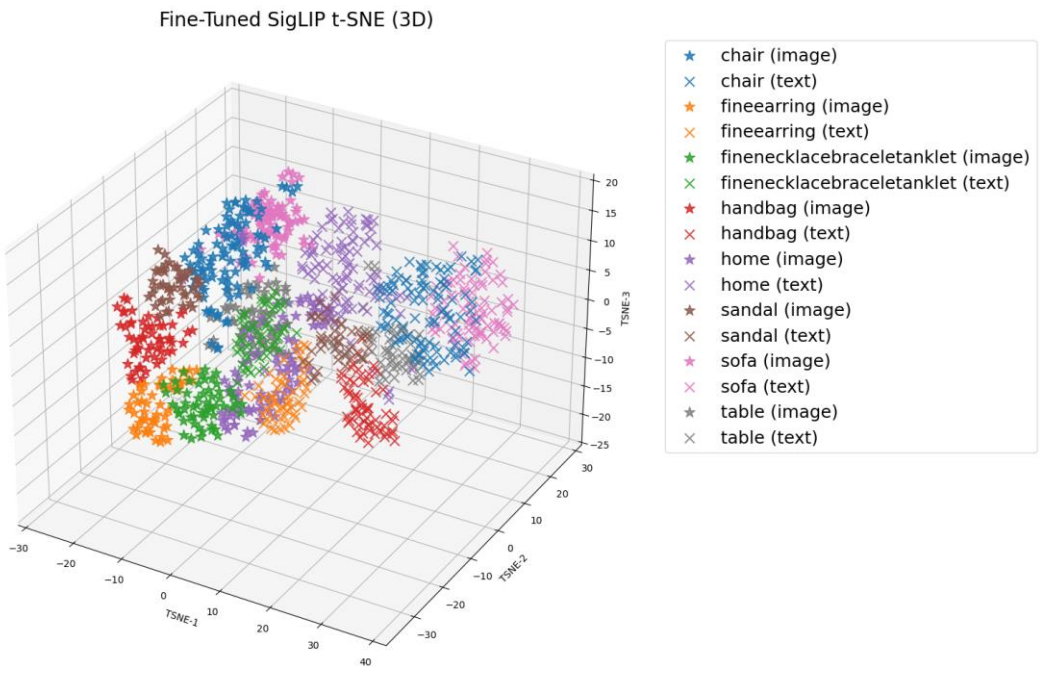

Figure 4: 3D t-SNE of fine-tuned SigLIP Embeddings for Eight Product Types

The 3D t-SNE plot in **Figure 3** shows the text and image embeddings before domain-specific fine-tuning. In this plot, the image embeddings (*) and text embeddings (×) for the same product type often form nearby but visibly separated sub-clusters, indicating a remaining modality gap—the model's representations for images and text are not yet well aligned for this specific ABO domain. At the same time, several product types appear compact and well clustered, suggesting the pretrained model already captures broad semantic structure. Some categories also partially overlap (e.g., furniture-related types such as *chair*, *sofa*, and *table*), which is expected because these items share visual and textual attributes.

Figure 4 shows the embeddings after fine-tuning on the curated ABO subset. Compared to **Figure 3**, the distance between text and image points within the same product type is generally reduced: text (×) and image (*) points are getting closer and more interleaved, which is consistent with improved cross-modal alignment on the target dataset. Additionally, the clusters appear more dispersed (larger spread) than in the pretrained plot, reflecting that fine-tuning encourages the model to separate individual items and fine-grained variations within each category rather than collapsing them into overly tight groups. Importantly, semantically similar categories such as *chair*, *sofa*, *table* still show some overlap, but the overall structure suggests the model is learning a more domain-appropriate embedding space where paired text–image representations are closer, while still preserving high-level semantic relationships across product types.

Table 1: Retrieval Performance Comparison Between Pretrained and Fine-Tuned Models

SigLip Model	Recall @ 1	Recall @ 3	MRR
Pretrained	0.62	0.81	0.73
Fine-tuned	0.77	0.95	0.86

After fine-tuning SigLIP ViT-B/16 on the curated ABO subset, retrieval performance showed clear and consistent improvements across multiple evaluation metrics as shown in **Table 1**.

- Recall@1, which measures the fraction of queries for which the correct product is ranked first, increased from approximately 0.62 with the pretrained model to 0.77 after fine-tuning. This indicates that the correct product now appears at the top 1 for roughly three-quarters of queries, a significant improvement over the baseline.
- Similarly, Recall@3 rose from 0.81 to 0.95, meaning that in most cases the correct product is now included among the top three retrieved products, enhancing the likelihood of a successful user search experience.
- Mean Reciprocal Rank (MRR) improved from 0.73 to 0.86, showing that correct matches not only occur more frequently but also tend to appear very near the top of the ranked results, which is crucial for fast and efficient product discovery.

Qualitative evaluation further supports these quantitative gains: text queries such as "modern gray fabric sofa" in **Figure 5** return products that closely match the described style, color, and category, while image queries as shown in **Figure 6** using uploaded product photos reliably retrieve visually and semantically similar items. These improvements demonstrate that the fine-tuned embeddings capture domain-specific characteristics of the ABO subset, aligning both images and text in a shared representation space that facilitates accurate and meaningful retrieval.

Figure 5: Search products by text queries

Figure 6: Search products by image queries

Overall, these results indicate that fine-tuning on the ABO subset significantly enhances the model's ability to match products across modalities, providing both higher accuracy and better user experience for multimodal product search.

Assessment of Project Outcomes

The project successfully achieved its main goal of building a stronger multimodal product search system for the subset of ABO catalog. Fine-tuning SigLIP ViT-B/16 led to clear quantitative gains: Recall@1 improved from about 0.62 to 0.77, Recall@3 from 0.81 to 0.95, and MRR from 0.73 to 0.86, indicating that the correct product is now found earlier and more consistently in the ranked results. Qualitative behavior supports these metrics: both text and image queries return visually and semantically appropriate products, and t-SNE visualizations show image and text embeddings from the same product category appear more dispersed yet moving closer together, indicating the improvement in text and image alignment. The combination of a fully fine-tuned dual encoder with FAISS-based similarity search provides a practical and scalable foundation for multimodal retrieval. Future work could explore real-time inference with low latency, expanding to additional product categories, and validating performance through large-scale online evaluation.

Challenges and Limitations

- **Class imbalance:** Strong class imbalance in the ABO subset (a few dominant categories, many rare ones), so performance is likely weaker on underrepresented product types.
- **Image coverage:** Most listings are represented primarily by a single main catalog image, and additional associated images are often less informative (e.g., close-up corners, diagrams, packaging, or accessory views). This reduces visual variety during training and can limit robustness to real-world user photos that differ in viewpoint, background clutter, lighting, and image quality.

- **Computational Expense:** Full fine-tuning of a large vision–language model is computationally expensive and sensitive to hyperparameters, requiring substantial GPU resources and careful tuning.

- **Scaling Complexity:** The current similarity search setup is validated at ~11k products; scaling to millions will need more advanced indexing and explicit latency–accuracy trade-off design.
- **Offline Evaluation:** Evaluation is purely offline with no personalization or online A/B testing, so real-world impact on user engagement and conversion is not yet measured.

Conclusion

This project demonstrates that fine-tuning a large vision–language model on a curated ABO subset can substantially enhance multimodal product search for e-commerce. By adapting SigLIP ViT-B/16 to the specific catalog and integrating it with FAISS index search, the system learns a shared embedding space where related images and texts are closely aligned, and category clusters are clearly separated. The observed improvements in retrieval metrics, along with the stable training behavior, indicate that the approach is effective at the current scale. Remaining limitations such as class imbalance, single-image coverage, and lack of live evaluation highlight clear opportunities for future work, including scaling to larger catalogs, supporting real-time inference with low latency, and incorporating personalization to enhance user experience and engagement.

References

Amazon. *Amazon Berkeley Objects Dataset (ABO)*. Amazon Web Services, n.d., https://registry.opendata.aws/amazon-berkeley-objects. Accessed 15 Oct. 2025.

Zhai, Xiaohua, et al. "Sigmoid Loss for Language Image Pre-Training." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language Supervision." *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-Scale Similarity Search with GPUs." *IEEE Transactions on Big Data*, vol. 7, no. 3, 2021, pp. 535–547.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.