Fine-Tuning OpenAI CLIP for Fashion Understanding: A Comparative Study with FashionCLIP

By
Amrutha Karuturi
NetID: ak2508

A research project submitted in partial fulfilment of the requirements of the Degree of Master of Science in Statistics Specialization in Data Science

at

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

May 2025

Abstract

This project explores the fine-tuning of OpenAI's CLIP model on a fashion dataset consisting of 44K images, aiming to improve its performance on fashion-specific tasks. The results will be compared with both the original CLIP model and the FashionCLIP model (which was pre-trained on 400K image-text pairs). The study evaluates the models on various downstream tasks, including image retrieval from text queries, zero-shot classification, and feature extraction of fashion attributes, where the output highlights specific regions of an image corresponding to the attributes mentioned in the query. The project aims to assess the effectiveness of fine-tuning CLIP for fashion applications and its potential for enhancing fashion-focused ML models.

Table of Contents

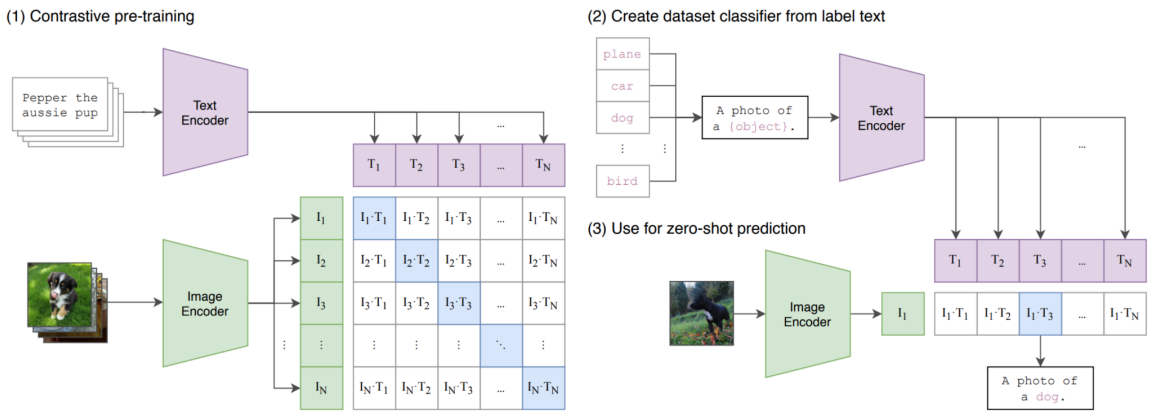
1.INTRODUCTION

Contrastive Language–Image Pretraining - CLIP (Radford et al., 2021), developed by OpenAI, is a powerful multimodal model that learns visual concepts from natural language supervision. Unlike traditional vision models that require manual labeling, CLIP is trained on a large dataset of image–text pairs using a contrastive loss. It jointly embeds images and their corresponding text descriptions into a shared feature space, allowing the model to measure the semantic similarity between an image and a text prompt. This enables CLIP to perform various vision-language tasks such as zero-shot classification, image-to-text and text-to-image retrieval, and visual grounding—without requiring task-specific fine-tuning.

The architecture of CLIP consists of two main components: an image encoder (typically a Vision Transformer or ResNet) and a text encoder (a Transformer-based model). During training, both encoders learn to map their respective modalities into a common embedding space where matching image–text pairs are pulled closer together, while non-matching pairs are pushed apart. Once trained, the model can evaluate how well any image matches a given text by computing cosine similarity in the shared space.

Despite CLIP's generalization capabilities, its performance can be suboptimal in specialized domains like fashion, where subtle visual details and domain-specific vocabulary are critical. This motivates the need to fine-tune CLIP on curated datasets tailored to fashion-related tasks. This project explores such a fine-tuning process using a 44,000-sample fashion dataset, evaluating how adaptation to this domain impacts performance on key downstream tasks such as image retrieval, zero-shot classification, and attribute-focused visual explanations.

Fig. 1 — Overview of CLIP Training and Zero-Shot Inference Workflow



2.DATA COLLECTION

The dataset used in this project is a structured fashion product dataset comprising 44,392 entries, each corresponding to a unique fashion product. Each entry includes an associated image and a set of descriptive attributes, making it well-suited for training and evaluating multimodal models like CLIP. The dataset is available in two variants: one with high-resolution images for detailed visual analysis, and another with lower-resolution images optimized for environments with limited computational resources.

Each product entry is accompanied by the following attributes:

- id: A unique identifier for each product.
- gender: Specifies the target audience, with values including Men, Women, Boys, Girls, and Unisex.
- masterCategory: A high-level category that broadly defines the product type. Examples include Apparel, Accessories, Footwear, Personal Care, Free Items, and Sporting Goods.
- subCategory: A more specific categorization within the master category. Examples include Topwear, Bottomwear, Watches, Bags, Socks, Lingerie and Nightwear, Sarees, and Shoes.
- articleType: The most fine-grained categorization, providing the specific product type. It includes terms such as T-shirts, Casual Shoes, Face Scrub, Shoelaces, Body Lotion, Cushion Covers, Hair Accessories, etc.
- baseColour: The primary color of the product (e.g., Black, Red, Blue).
- season: Indicates the recommended season for the product, such as Spring, Summer, Fall, or Winter.
- year: The year the product was released or manufactured.
- usage: Describes the intended use case for the product, such as Casual, Ethnic, Sports, Formal, or NA (not available).
- productDisplayName: A descriptive caption or name of the product, often used as a textual representation in image-text pairing.

This hierarchical structure—from master category to article type—offers rich semantic information that is essential for training domain-specific vision-language models. The diversity of product types and attribute combinations within the dataset also makes it a robust benchmark for evaluating the performance of fine-tuned CLIP models on fashion-specific tasks such as zero-shot classification, attribute recognition, and image retrieval

3.EXPERIMENTS

3.1 Methodology

The primary objective of the experimental setup was to explore strategies for fine-tuning OpenAI's original CLIP model on a domain-specific fashion dataset and assess whether its performance could be improved to approach that of the FashionCLIP model (Chia et al., 2022), which was pre-trained on a much larger fashion-oriented dataset (approximately 400,000 image-text pairs). In contrast, the current dataset contains only 44,000 image-caption pairs, which introduces a challenge of data scarcity for effective fine-tuning.

To evaluate the performance of CLIP models on fashion-specific tasks with limited training data, a series of experiments were conducted in multiple phases.

1. Data Preprocessing:

The fashion dataset included paired image-text samples, where each caption corresponded to a product image. Preprocessing involved:

- Text normalization: lowercasing and removal of special characters.
- Caption augmentation: combining product metadata (gender, master category, subcategory, article type, base color, and usage) to generate structured, descriptive captions.
- Image variants: both low-resolution and high-resolution images were used to analyze performance trade-offs under different computational constraints.
- Brand and Attribute Extraction via NER-based Modeling: As the dataset lacked a dedicated brand column, a custom NER model was trained on labeled data to extract structured entities from the product captions. Using BIO tagging, the model identified components such as gender, base color, brand, article type, and a broader product attribute category (e.g., "turtleneck," "slim fit"). This approach allowed for cleaner and more informative structured captions, enhancing alignment between text and image content.

2. Baseline Evaluation:

To establish performance bounds, the following baseline evaluations were performed:

- The **original CLIP model** was tested on the dataset for the image-text retrieval task, serving as a reference for how a general-purpose vision-language model performs on fashion-specific queries.
- The **FashionCLIP model**, pre-trained on a large-scale fashion dataset, was evaluated on the same retrieval task to provide an upper-bound benchmark.

3. Fine-Tuning with Raw Captions:

The CLIP model was first fine-tuned using the productDisplayName column from the dataset as the text input. Only basic text preprocessing (e.g., lowercasing and punctuation removal) was applied. This setup was used to assess the performance gain achievable with minimal effort in text augmentation.

- **Model Fine-Tuning Details**: The full CLIP model (both image and text encoders) was fine-tuned end-to-end using the original contrastive learning objective. No parts of the model were frozen, allowing all parameters to adapt to the fashion domain through supervised training on the paired image-caption data.

4. Fine-Tuning with Structured Captions:

To improve the quality of the text representation, structured captions were generated by concatenating multiple metadata fields (e.g., gender, article type, base color). The CLIP model was then fine-tuned using these enriched captions. This phase evaluated whether detailed, attribute-specific text could improve retrieval and classification performance.

5. Image Quality Comparison:

The dataset provides two sets of images—low-resolution and high-resolution. Both variants were used in separate experiments to examine the effect of image quality on model performance. While the performance difference was minimal for most retrieval tasks, certain cases, such as brand-based retrieval, showed slightly better results with high-resolution images.

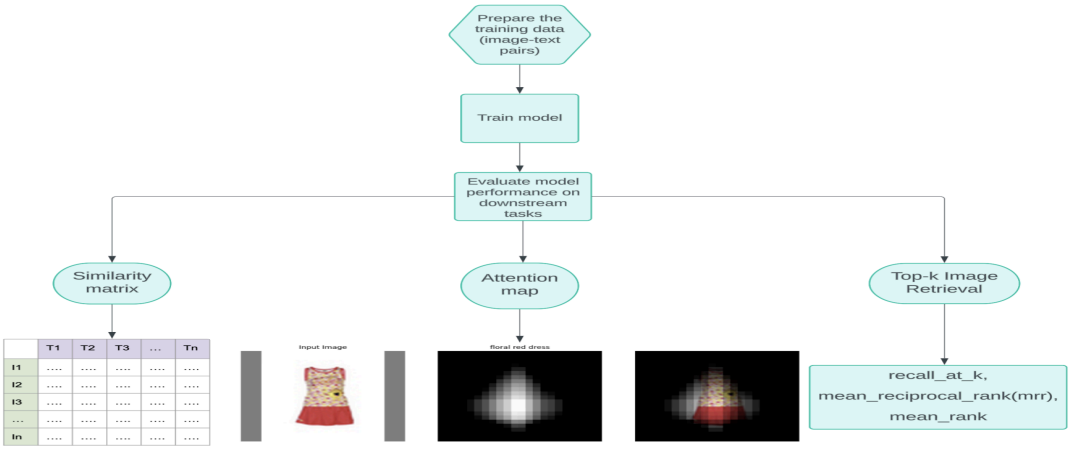
6. Evaluation Metrics:

Each model—baseline, FashionCLIP, and fine-tuned variants—was evaluated using the following methods:

- **Image–Text Similarity Matrix**: For a batch of image-text pairs, cosine similarity scores were computed between the image and text embeddings. The diagonal elements of the similarity matrix (representing correct matches) were expected to have the highest values. Improvements in diagonal dominance across models indicated better alignment between matching pairs.
- **Top-K Retrieval Accuracy:**
  - Recall@K: Measures the proportion of queries for which the correct image appears in the top-K retrieved results. Higher Recall@K indicates that the model is more likely to retrieve relevant matches within the top K results.
  - Mean Reciprocal Rank (MRR): Computes the average of reciprocal ranks of the correct image across all queries, where reciprocal rank is 1/rank. Higher MRR means the correct image tends to appear closer to the top of the retrieval list.
  - Mean Rank (MR): Calculates the average rank position of the correct image across all queries. Lower Mean Rank indicates better performance, as correct images appear earlier in the retrieval list.
- **Attention Heatmaps**: Visual explanations were generated to understand which parts of an image the model focused on when attending to specific queries (e.g., "shoes," "watch"). This helped assess whether the fine-tuned models improved their ability to localize relevant attributes in fashion images.

Through these controlled experiments, the study aims to quantify the contribution of text enhancements and image quality to the fine-tuning process and identify the most effective setup for improving the general CLIP model's fashion-specific capabilities.

Fig. 2 — Overview of project workflow



3.2 Base Models and Results:

To evaluate the effectiveness of pre-training on fashion-specific data, we compared the original CLIP model and the pre-trained FashionCLIP model across three evaluation dimensions: image–text similarity matrix, attention heatmaps, and top-K retrieval performance

.

1. Similarity Matrix Evaluation

For a random sample of five image–text pairs, cosine similarity matrices were computed between image embeddings and their corresponding caption embeddings. Ideally, the diagonal entries in the matrix (representing correct image–caption pairs) should have the highest values.

- CLIP Model: Diagonal values were generally higher than the off-diagonal entries, indicating correct matches, but the margin between matching and non-matching pairs was modest.
- FashionCLIP Model: Demonstrated significantly better separation between correct and incorrect pairs. Non-diagonal values were much more suppressed, for example: 0.177 → 0.065, 0.2966 → 0.11. Meanwhile, diagonal values increased noticeably, suggesting stronger and more confident associations between images and their corresponding captions.

Fig.3 – Similarity matrix of embeddings from CLIP(left) and FashionCLIP(right) models

	T1	T2	T3	T4	T5
I1	0.225	0.114	0.191	0.162	0.174
I2	0.169	0.303	0.138	0.216	0.175
I3	0.246	0.177	0.330	0.206	0.229
I4	0.254	0.258	0.210	0.304	0.208
I5	0.162	0.145	0.179	0.173	0.308

	T1	T2	T3	T4	T5
I1	0.336	0.065	0.129	0.124	0.081
I2	0.039	0.355	0.005	0.133	0.042
I3	0.211	0.030	0.353	0.064	0.171
I4	0.114	0.201	0.112	0.331	0.084
I5	0.086	0.044	0.116	0.045	0.316

2. Top-K Retrieval Performance

To evaluate quantitative retrieval accuracy, metrics such as Recall@K, Mean Reciprocal Rank (MRR), and Mean Rank (MR) were computed over a batch of query–image pairs:

- FashionCLIP consistently outperformed the CLIP model across all retrieval metrics.
- Notably, the Recall@1 and MRR values nearly doubled, indicating that FashionCLIP is significantly better at ranking the correct image near the top.
- The mean rank was also reduced by over 50%, showing that correct images were retrieved much earlier in the ranked list.

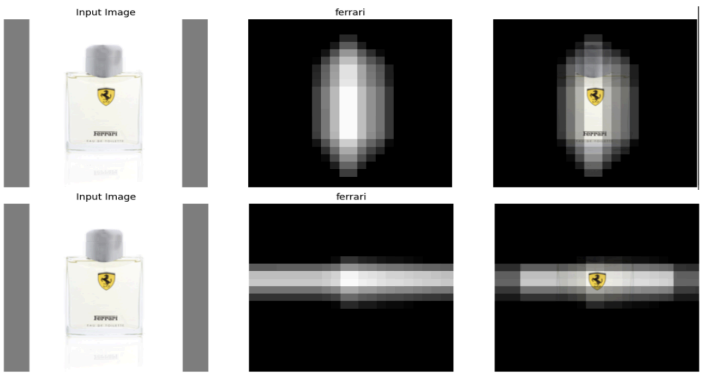


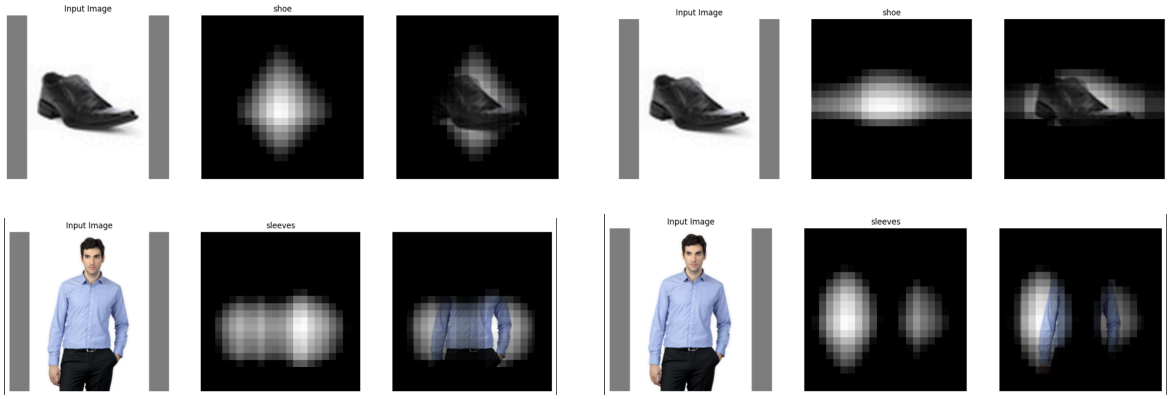
Fig.4 – Attention maps by CLIP(top) and FashionCLIP(bottom)

3. Attention Heatmaps

Visual attention maps were generated to understand how each model interprets the visual content in response to a specific text query. Several examples highlight the qualitative differences:

- Ferrari Perfume Bottle (Query: “Ferrari”):
  - CLIP Model: Focused on the Ferrari logo but also spread attention unnecessarily to horizontal background whitespace.
  - FashionCLIP: Accurately highlighted both the Ferrari logo and text on the bottle, with attention tightly focused on the object, effectively ignoring the background.
- Shoe Image (Query: “Shoe”):
  - CLIP Model: Highlighted the shoe reasonably well but in a somewhat diamond-shaped region that spilled into irrelevant background.
  - FashionCLIP: Formed a tighter, rhombus-shaped map that more precisely covered the shoe, showing greater spatial focus and contextual understanding.
- Person Wearing Long-Sleeved Shirt (Query: “Sleeves”):
  - CLIP Model: Successfully responded to the query but included parts of the torso in its attention.
  - FashionCLIP: Focused specifically on the sleeves, showing refined localization and better grasp of fashion-relevant semantics.

Fig. 5 – Examples for attention maps by CLIP(left) and FashionCLIP(right)



These results underscore the advantage of domain-specific pretraining for fine-grained tasks such as fashion item retrieval and attribute localization. FashionCLIP exhibits both quantitatively higher retrieval accuracy and qualitatively better visual grounding within fashion contexts. The baseline evaluations were designed not only to highlight the performance gap between general-purpose and domain-specific models but also to establish reference points for evaluating the effectiveness of the custom fine-tuned CLIP model. The original CLIP model, trained on general data, serves as a lower bound, illustrating the limitations of a generic vision-language model on fashion-specific tasks. In contrast, FashionCLIP, pretrained on 400K fashion-centric image–text pairs, represents a strong upper bound. By positioning the custom fine-tuned model’s performance between these two baselines, we can assess how well targeted fine-tuning, even with significantly less data, bridges the gap toward domain adaptation and specialization.

3.3 Fine-tuned models

To enhance CLIP's performance on fashion-specific tasks, several fine-tuning strategies were applied using the 44K image–text pairs. Each strategy focused on improving the image-text alignment through caption restructuring, image quality variations, and metadata extraction.

1. Metadata-Enriched Captions*(model2)*:

The original dataset included a productDisplayName field as a free-text caption. However, it often lacked important product attributes such as gender, season, or usage. To provide a richer textual description, structured captions were generated by appending metadata fields—only if they were not already present in the original caption. For example, if the base color or gender was missing in the caption, it was added from the respective column. This ensured that the model received more complete information without introducing redundancy. The goal was to make implicit product details explicit, improving the model's understanding and retrieval accuracy.

2. Image Quality Variants:

To evaluate the impact of image resolution on model performance:

- Low-resolution images were used for initial fine-tuning due to their reduced computational cost, enabling faster experimentation*(model1)*.
- High-resolution images were later used to assess if finer visual details could improve attribute-level recognition, especially in cases like brand logos or intricate design elements*(model3)*.

Early results showed marginal improvement with high-resolution inputs. However, brand-specific retrieval—such as identifying the "Ferrari" label on a perfume bottle—showed the potential benefit of detailed imagery. For instance, FashionCLIP successfully highlighted both the logo and the "Ferrari" text using attention maps, while the original CLIP model could only focus on the symbol. This suggests that richer textual alignment and higher-resolution images might improve brand-level recognition in fine-tuned models.

3. NER-Based Attribute Extraction for Structured Captions*(model4)*:

To address the absence of explicit columns for brand names and product-specific attributes (e.g., "turtleneck," "slim fit"), a custom Named Entity Recognition (NER) model was trained to extract these entities from raw captions.

- The task was framed as a multi-label sequence tagging problem using BIO (Begin-Inside-Outside) format.
- The model was trained on labeled caption data(*Fig. 6*) with five categories: Gender, Base Colour, Brand, Article Type, and a broader Product Attribute class.
- Instead of generating new words, the model tagged existing tokens in the captions that belonged to these categories(*Fig. 7*).

This allowed automatic transformation of unstructured captions into consistent, structured templates, which improved model focus during training. Rather than requiring CLIP to infer relationships across varied sentence structures, the structured format provided a clearer mapping between text and visual concepts, enhancing retrieval and attention performance.

Fig. 6 - Input text tokenization and label assignment for training the NER model

- **Caption**: Catwalk Women Gun Metal Grey Heels
- **Tokens**: ['cat', 'walk', 'woman', 'gun', 'metal', 'grey', 'heel']
- **Labels**: ['B-BRAND', 'I-BRAND', 'B-GENDER', 'B-ATTR', 'I-ATTR', 'B-COLOR', 'B-ARTICLETYPE']

Fig. 7 - Example of named entities detected by the fine-tuned NER model

- **Caption**: `"hanes grey melange shorts for men"`
- **Model Output**:

  {entity_group: `BRAND`, score: `0.99975365`, word: `hanes`, start: 0, end: 5}, {entity_group: `COLOR`, score: `0.99123037`, word: `grey`, start: 6, end: 10}, {entity_group: `ATTR`, score: `0.9916775`, word: `melange`, start: 11, end: 18}, {entity_group: `ARTICLETYPE`, score: `0.7740248`, word: `shorts`, start: 19, end: 25}, {entity_group: `GENDER`, score: `0.9972223`, word: `men`, start: 30, end: 33}

Fig. 8 - Similarity matrix of embeddings from model2, model1

	T1	T2	T3	T4	T5
I1	0.312	0.129	0.126	0.098	0.092
I2	0.114	0.328	0.116	0.191	0.126
I3	0.210	0.126	0.311	0.127	0.165
I4	0.099	0.216	0.100	0.357	0.104
I5	0.088	0.106	0.176	0.090	0.315

	T1	T2	T3	T4	T5
I1	0.300	0.117	0.118	0.123	0.093
I2	0.087	0.297	0.056	0.119	0.090
I3	0.199	0.108	0.309	0.121	0.145
I4	0.087	0.179	0.104	0.320	0.088
I5	0.079	0.050	0.127	0.058	0.280

Fig. 9- Similarity matrix of embedding from model3, model4

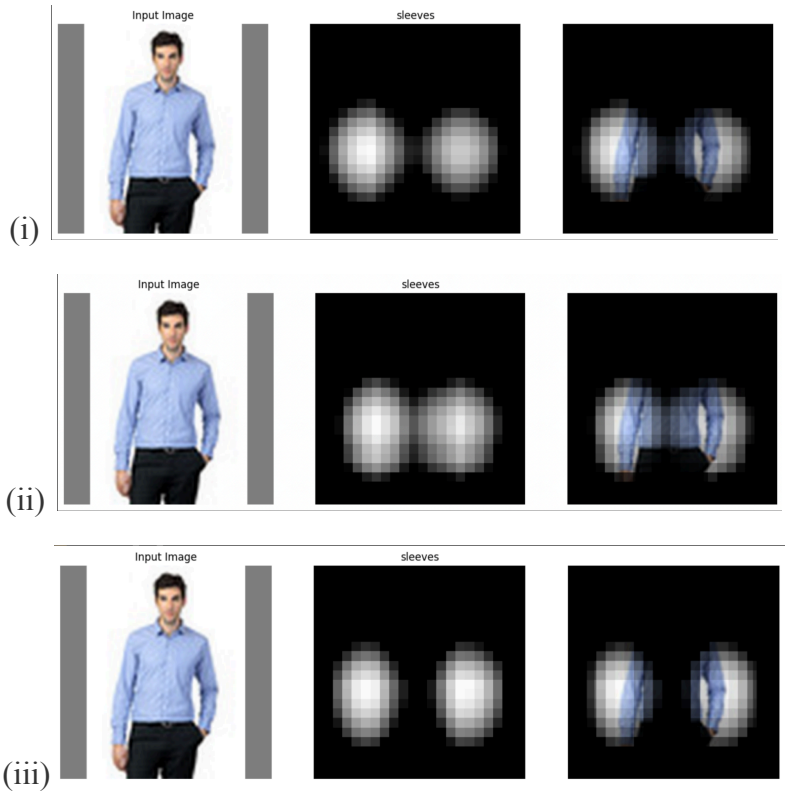
	T1	T2	T3	T4	T5
I1	0.354	0.050	0.117	0.056	0.071
I2	0.041	0.309	0.074	0.170	0.094
I3	0.155	0.128	0.315	0.103	0.121
I4	0.033	0.176	0.114	0.332	0.086
I5	0.070	0.082	0.148	0.052	0.313

	T1	T2	T3	T4	T5
I1	0.331	0.099	0.121	0.127	0.100
I2	0.073	0.300	0.077	0.159	0.084
I3	0.174	0.144	0.301	0.140	0.153
I4	0.092	0.204	0.099	0.341	0.093
I5	0.140	0.088	0.160	0.147	0.298

Fig. 10 - Performance metrics of all models

Model	**Recall@1**	**Recall@5**	**Recall@10**	**MRR**	**Mean Rank**
CLIP Model	0.1690	0.4010	0.5355	0.2851	40.56
Fashion CLIP Model	0.2917	0.5898	0.7292	0.4321	18.03
Model 2	0.2784	0.6166	0.7662	0.4337	12.12
Model 3	0.2773	0.6179	0.7740	0.4334	11.55
Model 4	0.2255	0.5510	0.7105	0.3749	14.48

Fig. 11 - Attention maps by model2 ,model3 , model4



4. RESULTS AND OBSERVATIONS

Key Observations from the Similarity Matrices(*Fig. 3, 8, 9*):

- **Baseline Comparison (Original CLIP vs. FashionCLIP):** FashionCLIP exhibits sharper diagonal dominance compared to the original CLIP, indicating stronger alignment between correct image-text pairs. CLIP's embeddings remain more diffused, reflecting its general-purpose training.
- **Model1: Fine-tuning on Direct Captions**: No significant improvement was observed in diagonal matching scores. However, a noticeable **reduction in similarity for negative pairs** (off-diagonal entries) was observed, indicating improved discrimination between unrelated image-text pairs.
- **Model2: Fine-tuning with Crafted Captions (Augmented)**: Both **diagonal matching scores increased** and **negative pair similarities decreased**, leading to better separation between correct and incorrect matches. Among the models, this setting showed one of the clearest improvements in retrieval alignment.
- **Model3: Fine-tuning with High-Resolution Images**: Slight localized improvements were observed in diagonal scores for certain samples. However, **overall improvements were limited**, and no strong reduction in negative similarities was observed.
- **Model4: Fine-tuning with Structured Captions (NER-based Attributes)**: **Strong improvement in diagonal matching scores** was observed, showing that structured, attribute-focused captions help CLIP models better specialize. Additionally, negative pair similarities were suppressed, making Model4 the most consistently improved fine-tuned model among all experiments.

Fine-tuning strategies that enriched the textual input (Model2 and Model4) achieved the most effective improvements, either by boosting correct match scores or reducing incorrect match similarities. Training only on higher-resolution images (Model3) showed comparatively smaller benefits under resource constraints.

Key Observations from the Retrieval metrics(*Fig. 10*):

- **CLIP Model**: The original CLIP model shows the weakest retrieval performance, meaning it often fails to retrieve the correct image among the top matches and retrieves correct items only after many wrong ones (high Mean Rank).
- **FashionCLIP Model**: FashionCLIP achieves higher Recall@1 and a much lower Mean Rank, meaning it retrieves the correct product near the top much more reliably thanks to fashion-specific pretraining.
- **Model 1 (Fine-tuned on Direct Captions)**: Fine-tuning on raw captions leads to a strong jump in performance, with higher Recall@1 and a high MRR, meaning the model retrieves the correct image among the top results more frequently.
- **Model 3 (Fine-tuned with High-Resolution Images)**: Using high-resolution images slightly improves retrieval sharpness, reflected in the best Mean Rank among fine-tuned models, meaning it needs fewer retrieval steps on average to find the correct image.
- **Model 4 (Fine-tuned with Structured Captions via NER)**: Although Recall@1 is slightly lower, the model maintains strong Recall@10, meaning it places the correct result within the top 10 results consistently, which is important for broader search and recommendation settings.

Key Observations from the Attention map(*Fig. 5, 11*):

- **Original CLIP**: Broad and diffused attention with poor localization on the target attributes, often spreading across irrelevant regions.
- **FashionCLIP**: Focused and sharper attention exactly over relevant regions (e.g., shoes, sleeves), showing better specialization for fashion concepts.
- **Model2 (crafted captions)**: Attention improves noticeably, with better focus around target areas like sleeves, although still somewhat coarse compared to FashionCLIP.
- **Model3 (high-resolution images)**: Attention regions become sharper and more detailed, but without significant improvement in the exact localization compared to Model2.
- **Model4 (structured captions via NER)**: Most precise and clean attention among fine-tuned models, accurately highlighting the sleeves with minimal background distraction.

The trends observed in the similarity matrices are consistent with the qualitative insights from the attention maps. Models that showed better separation of positive and negative pairs (such as Model2 and Model4) also produced more focused and accurate attention regions corresponding to the queried attributes. In particular, fine-tuning with structured or enriched captions not only increased the correct retrieval scores but also helped the model localize visual features more precisely, as evidenced by sharper, attribute-specific activations around the sleeves. Conversely, models with limited improvement in similarity separation (such as Model1) exhibited broader and less targeted attention distributions. Thus, the improvements in retrieval performance and image-text alignment metrics are clearly reflected in the visual interpretability of the models.

5. CHALLENGES & FUTURE DIRECTIONS

This study focused on fine-tuning OpenAI's CLIP model on a domain-specific fashion dataset to bridge the performance gap with FashionCLIP.
A major theme explored was how **different caption enrichment strategies** impact the model's ability to specialize when only a **limited number of image-text pairs** are available — a common scenario in domain adaptation tasks.

Several challenges were encountered:

- **Data Scarcity**: With only 44,000 training pairs, the model's specialization remained limited compared to FashionCLIP, which was trained on a much larger dataset.
- **Caption Limitations**: Raw product captions often lacked key descriptive attributes, making **caption enrichment** — through metadata combination and NER-based structured caption generation — a critical factor in improving alignment and retrieval performance.
- **Computational Constraints**: Due to resource limitations, experiments with high-resolution images were limited. In such settings, focusing on **text-side enrichment** proved more effective and scalable than relying solely on improving image inputs.

Among all approaches, **caption enrichment showed the most substantial impact**, helping the model better align text and image modalities and improving retrieval metrics even without access to massive datasets.

Future Directions:

- **Scaling Caption Enrichment**: Expanding attribute extraction and using external knowledge sources can further enhance the information content in captions.
- **Partial Freezing During Fine-tuning**: Freezing lower layers and fine-tuning only the higher (domain-adaptive) layers can improve training stability and efficiency when limited data is available.
- **Enhanced Attention Supervision**:Introducing targeted supervision for attention mechanisms could sharpen visual-text alignment and improve model interpretability.

6. CONCLUSIONS

Fine-tuning domain-agnostic vision-language models like CLIP can be made highly effective with relatively small datasets, provided that **caption enrichment is systematically applied**.

Given the common challenge of data scarcity in real-world domains, this work highlights that **smart text engineering combined** with selective model adaptation strategies can significantly boost specialization without requiring massive retraining or extensive computational resources.

7. REFERENCES

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.
2. Chia, P. J., Attanasio, G., Bianchi, F., Terragni, S., Magalhaes, A. R., Goncalves, D., ... & Tagliabue, J. (2022). Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, *12*(1), 18958.
3. Fashion Product Images Dataset (High resolution images) - https://www.kaggle.com/paramaggarwal/fashion-product-images-dataset
4. Fashion Product Images Dataset small (Low resolution images) - https://www.kaggle.com/paramaggarwal/fashion-product-images-small
5. Code for the project: Final project CLIP