AI-Driven Fashion Cataloging: Transforming Images into Textual Descriptions

Stanford CS224N Custom Project

SiYi Ma
Department of Computer Science
Stanford University
siyima00@stanford.edu

Nishant Gopinath
Department of Computer Science
Stanford University
nishgop@stanford.edu

Abstract

We designed our project to harness cogVLM's capabilities for transforming visual fashion imagery into precise, fashion-specific terminologies. It leverages advanced multi-modal learning to bridge the gap between visual perception and textual description. By fine-tuning cogVLM in conjunction with RoBERTa, our system aims to accurately identify and classify key fashion attributes (e.g., category, neckline, pattern) directly from images. This initiative seeks to automate and refine the process of generating rich, detailed product catalogs for the fashion industry, enhancing the granularity and utility of product metadata. Our work is poised to contribute significant advancements in how e-commerce platforms catalog fashion items, making product discovery more intuitive and aligned with specific consumer preferences.

1 Key Information

- Mentor: Yuhui Zhang
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

The rapid growth of the online retail market underlines the demand for more accurate and efficient methods of cataloging and describing fashion products. Traditional methods often fail to transform visual data into detailed and contextually rich textual descriptions. Current methods frequently need help to effectively bridge the gap between visual perception and textual description, hindering the automation of product catalog generation and the seamless shopping experience. Our project addresses this challenge by leveraging the power of multimodal artificial intelligence, specifically developing and fine-tuning a system based on cogVLM[1]. By employing cogVLM, we aim to narrow this gap significantly. cogVLM innovatively combines a frozen pre-trained language model with an image encoder, enhanced by a trainable visual expert module [1]. This approach promises to revolutionize how fashion products are cataloged online, demonstrating the practical application of multimodal AI in boosting e-commerce efficiency and meeting the evolving needs of the fashion online retail market.

This initiative will showcase the practical application of multimodal AI in seamlessly integrating visual data and natural language processing. Our efforts aim to offer a sophisticated solution that resonates with the growing demands of the online retail market in fashion. We strive to surmount the prevalent challenges in automated product catalog generation and establish a new standard for online presentation and discovery of fashion products.

3 Related Work

The CogVLM model is a powerful open-source visual language foundation model. It differs from the popular shallow alignment method, which maps image features into the input space of the language model. Instead, CogVLM bridges the gap between the frozen pretrained language model and image encoder by a trainable visual expert module in the attention and FFN layers[1]. This enables deep fusion of vision language features without sacrificing any performance on NLP tasks[1]. CogVLM-17B, a variant of CogVLM, has demonstrated state-of-the-art performance on 10 classic cross-modal benchmarks, including NoCaps, Flicker30k captioning, RefCOCO, RefCOCO+, RefCOCOg, Visual7W, and ScienceQA[1].

Another related model is CogAgent, an image understanding model developed based on CogVLM[2]. CogAgent is an open-source visual language model improved based on CogVLM. CogAgent-18B has 11 billion visual parameters and 7 billion language parameters, supporting image understanding at a resolution of 1120*1120. On top of the capabilities of CogVLM, it further possesses GUI image Agent capabilitieshong2023cogagent.

In the context of evaluating the robustness of large multimodal models (LMMs)[3], have introduced a comprehensive benchmark, named MMCBench, that covers more than 100 popular LMMs2. They specifically examined the self-consistency of their outputs when subjected to common corruptions[3].

Image-to-text models output a text from a given image. The most common applications of image to text are image captioning and optical character recognition (OCR)[4]. Image captioning is the process of generating a textual description of an image, which can help visually impaired people to understand what's happening in their surroundings[4]. OCR models convert the text present in an image, e.g., a scanned document, to text[4]. Pix2Struct is a state-of-the-art model built and released by Google AI. These tasks include captioning UI components, images including text, visual questioning infographics, charts, scientific diagrams, and more[4].

4 Approach

Integrated cogVLM and RoBERTa Framework: Our solution leverages the synergistic capabilities of cogVLM and RoBERTa to create an AI-driven fashion cataloging system. cogVLM, serving as the core, utilizes its advanced vision-language pre-training to decipher and extract nuanced features from fashion imagery. This allows us to understand the visual content at a granular level, accurately identifying and categorizing fashion attributes.

Enhanced Similarity Search with RoBERTa: We harness RoBERTa for its exceptional ability to perform similarity searches within fashion terminology. This is critical for matching the attributes extracted by cogVLM with accurate and relevant textual descriptions. RoBERTa is fine-tuned with cogVLM to ensure superior performance, focusing on refining the system's capacity to generate precise and contextually appropriate fashion descriptors.

Enhanced Similarity Search with RoBERTa: The efficacy of our approach, both before and after the application of model enhancements, is rigorously assessed using a custom, manually labeled dataset. This dataset is a benchmark, allowing us to measure improvements and fine-tune our models with high precision.

Collaborative Multi-Modal Learning: Our methodology embodies a collaborative fusion of cogVLM and RoBERTa, orchestrated specifically for fashion cataloging. This tailored multi-modal learning approach is pivotal, enabling us to interpret and analyze fashion content through a unique lens, markedly advancing the domain with our pioneering contributions.

Original Contributions: In addition to leveraging pre-trained models (cogVLM[1] and RoBERTa[2]) and fine-tuning demonstrations, all developed software, including Gradio apps for model inference, inference workflows utilizing diverse prompts and alternative multi-models, customized similarity search, fine-tuning code adaptation, bespoke evaluation metrics, and meticulous manual data labeling, aimed at comprehensively capturing fashion cataloging nuances.

[1] `https://github.com/THUDM/CogVLM`

[2] `https://huggingface.co/sentence-transformers/all-roberta-large-v1`

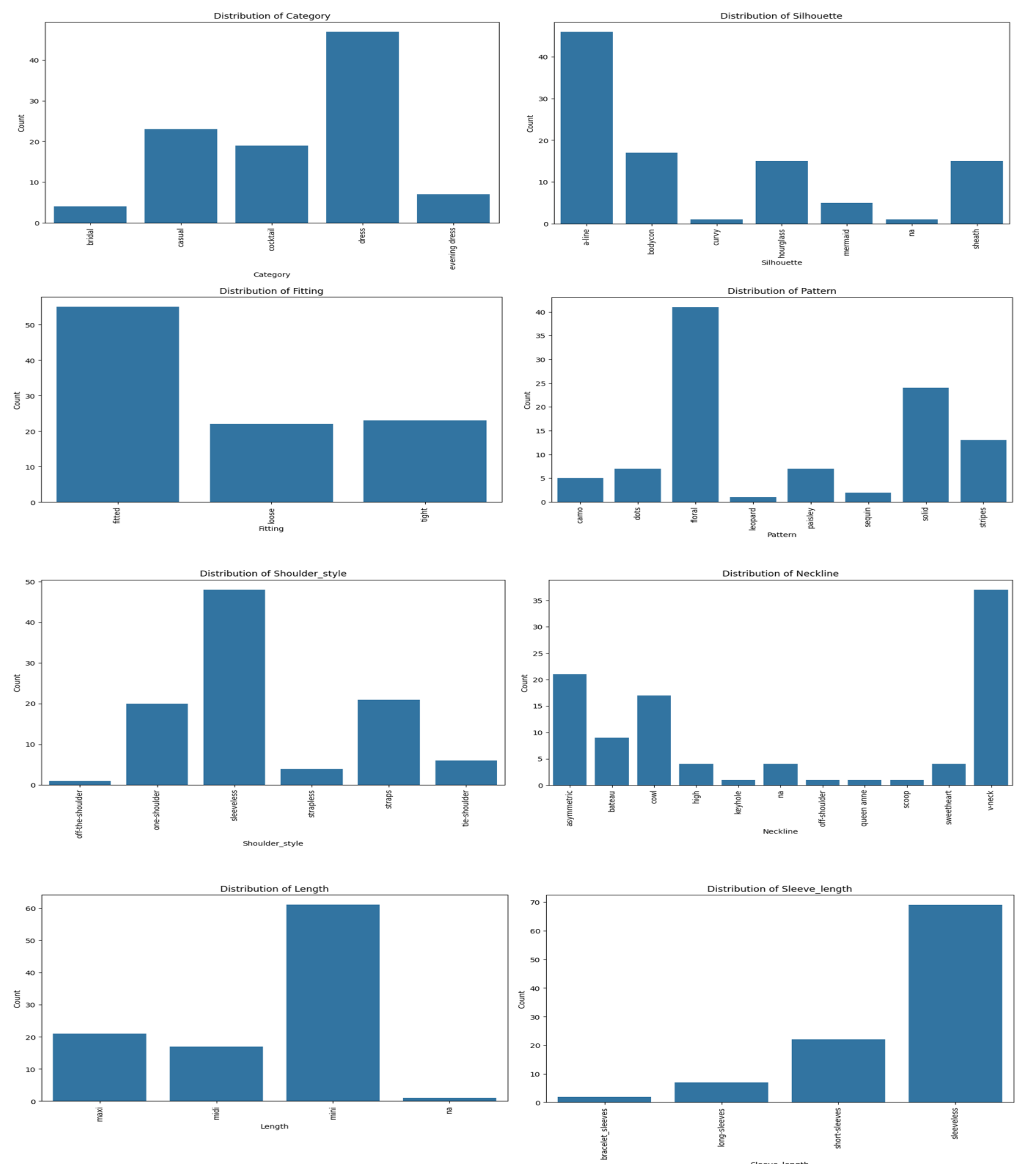
5 Data

5.1 Data

We make use of the Fashionpedia [5] dataset, which offers an extensive collection of fashion photos annotated for a range of attributes and categories. With 48,000 photos of regular people and celebrities wearing different looks, the Fashionpedia dataset offers thorough segmentation for apparel. The Fashionpedia dataset serves as a perfect reference point for training and evaluating our model.

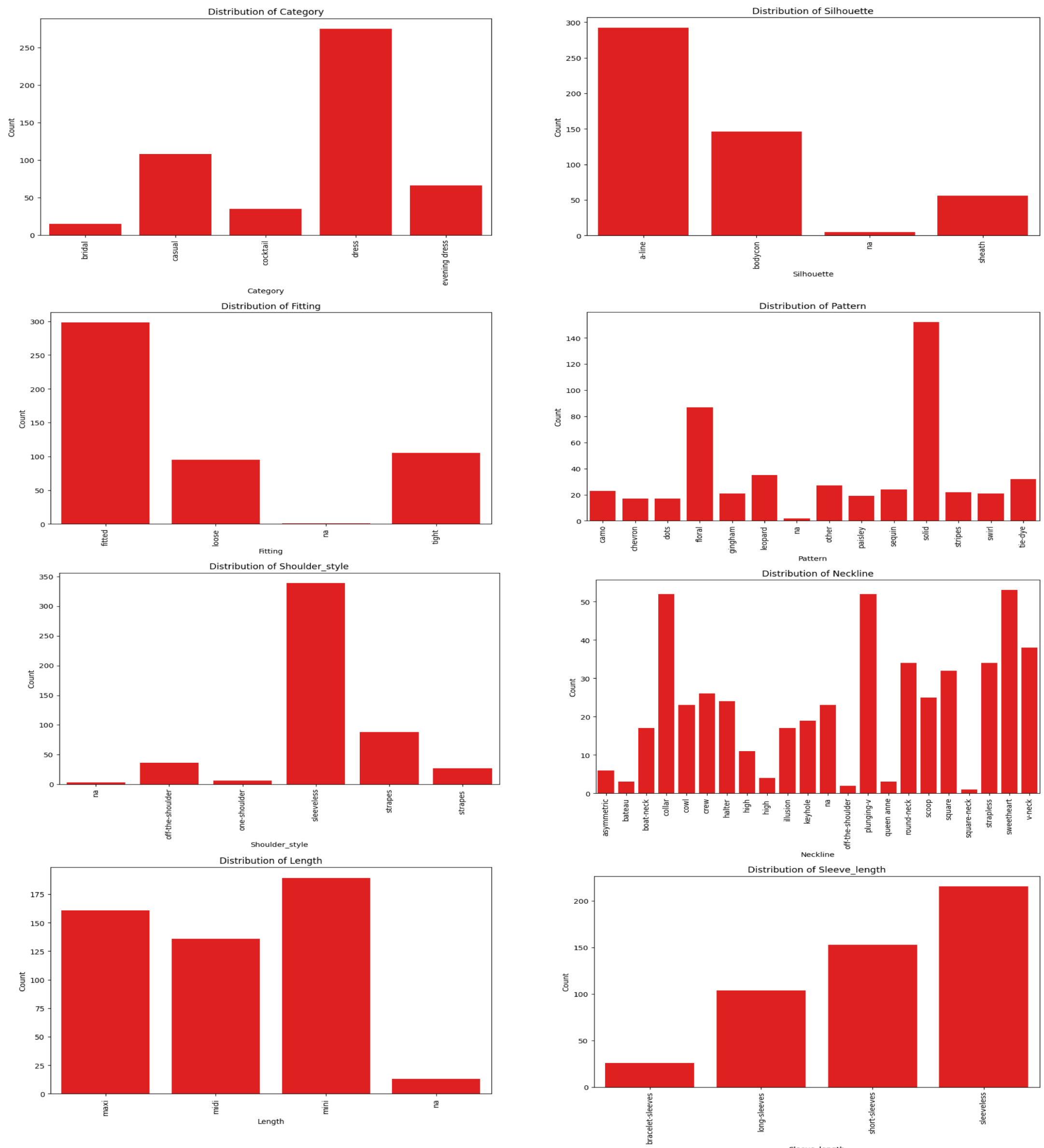
5.1.1 Test Data

We manually labeled a representative subset of the dataset (i.e., 100 images) to serve as ground truth. We then matched each manually labeled sample with its closest counterpart in the **cogVLM**-produced dataset based on semantic similarity by using **RoBERTa**-based similarity search to find the closest matches in the dataset generated by **cogVLM**.



5.1.2 Training Data

To enhance the performance of our model, we manually labeled an additional 500 images to create a fine-tuning dataset. Our goal in fine-tuning the model on this larger dataset was to improve its accuracy and performance in attribute extraction tasks. Additionally, we utilized the initial 100 manually labeled images to generate results, both before and after fine-tuning the model. By comparing the two, we can determine how fine-tuning affects the model's ability to extract attributes and evaluate the improvements made during the process.

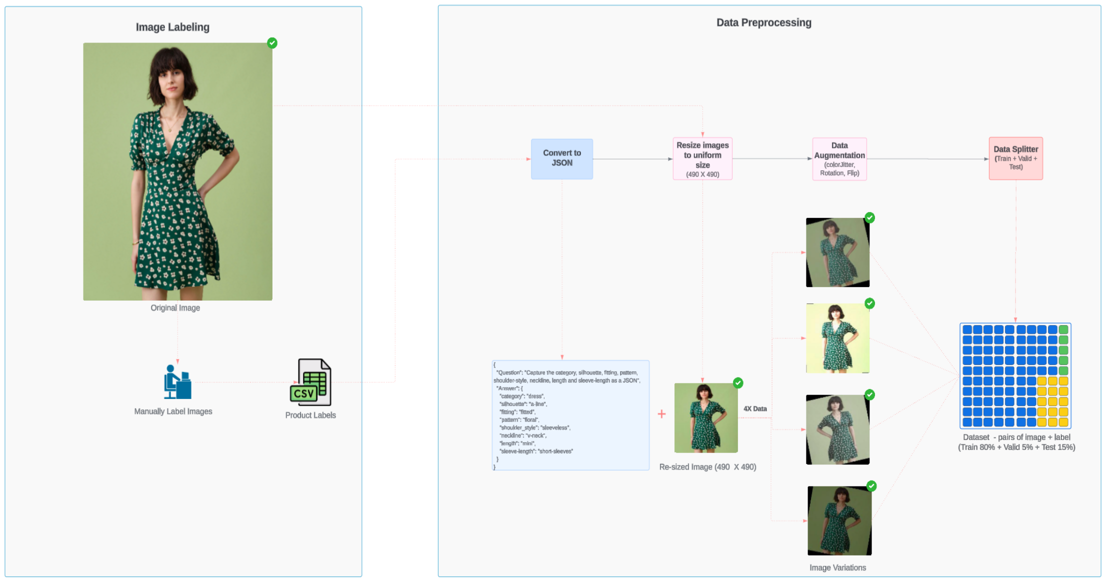


5.1.3 Data Preparation for Fine-tuning:

- Conversion from CSV to JSON: We convert the product labels from CSV format to JSON to streamline integration with subsequent steps or to comply with the input requirements of the machine learning framework.

- Resize Images to 490 x 490: We resize all images to a uniform size of 490x490 pixels. This standardization is essential for achieving the consistency neural networks require to process the images efficiently.

- Data Augmentation: The image dataset is augmented, likely through methods such as color jitter, rotation, flip, and others, to increase the dataset's size and variability, which can help improve the model's generalization and robustness.
- Data Splitting (Train, Valid, Test): We split the augmented dataset into training, validation, and testing. This is a standard practice for training the model, tuning hyperparameters, and finally evaluating the model's performance on unseen data.



6 Experiments

6.1 Evaluation method

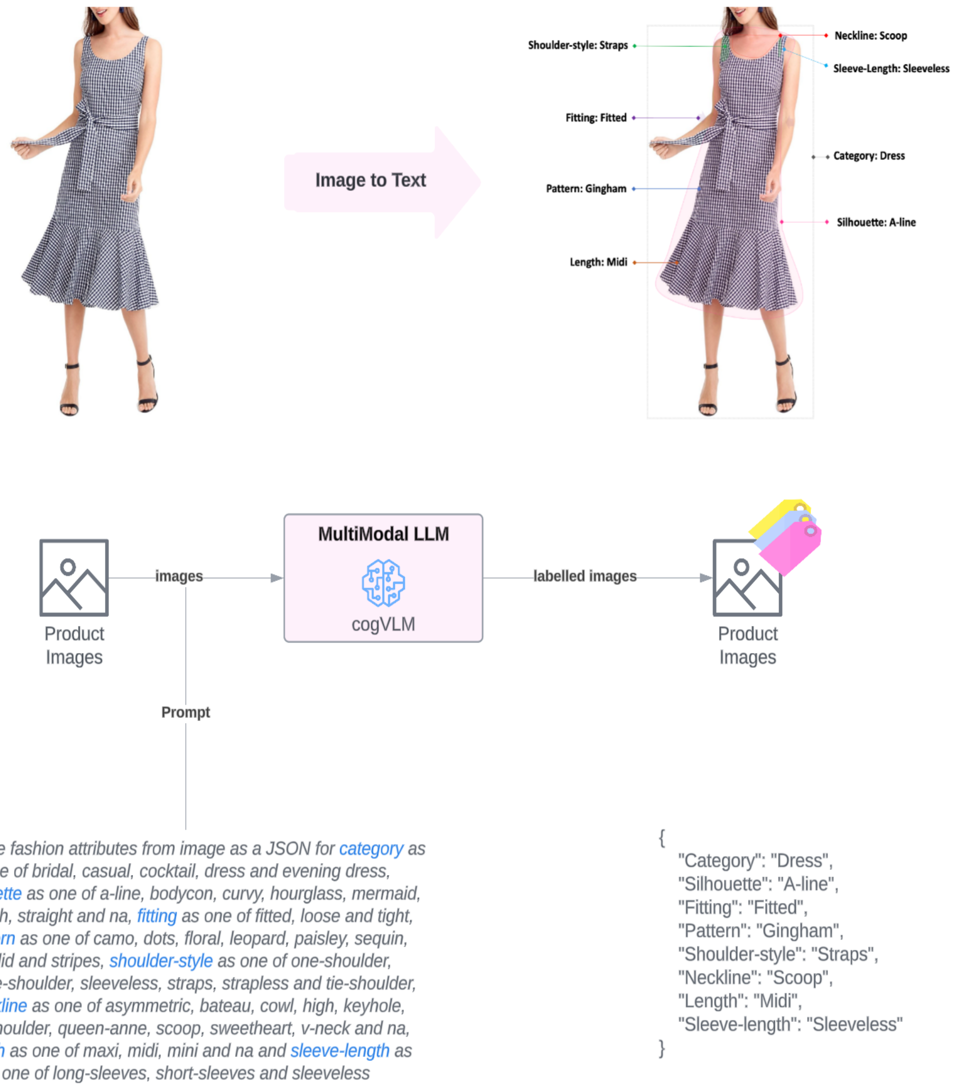
Our methodological framework involves selectively fine-tuning **cogVLM** using a curated subset of women's dresses from the Fashionpedia [5] dataset, enriched with manually labeled data to ensure comprehensive coverage of fashion-specific attributes, thus enhancing the model's training. We also use a **RoBERTa**, a sentence-transformer, to find similarity of the model-inferred label values to manually defined label values to uniquely created tags for distinct fashion attributes such as category, silhouette, fitting, pattern, shoulder-style, neckline, length, and sleeve-length. This tailored approach aims to enhance the model's ability to identify and classify these attributes precisely, leveraging **cogVLM**'s visual-language understanding and manual annotations.

Furthermore, we conduct comparisons of **cogVLM** against popular alternate multi-modal models such as **GPT4-Vision**[3], **Llava 1.5**[4], and **Qwen-VL**[5], constructing separate inference workflows for each.

[3] https://platform.openai.com/docs/guides/vision
[4] https://huggingface.co/liuhaotian/llava-v1.5-7b
[5] https://huggingface.co/Qwen/Qwen-VL-Chat

6.2 Experimental details

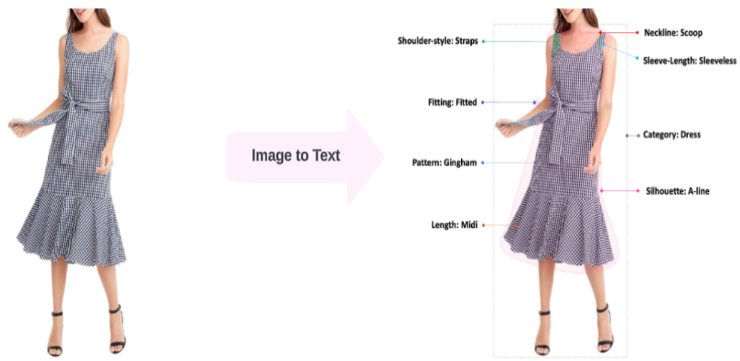
Approach A: Prompt with possible value list



Description: A predefined prompt extracts image attributes with a specified possible value list in this approach. The prompt structure captures each attribute (or label class) along with the attribute values (label values), prompting the model to describe the image in terms of the defined label class and label values.

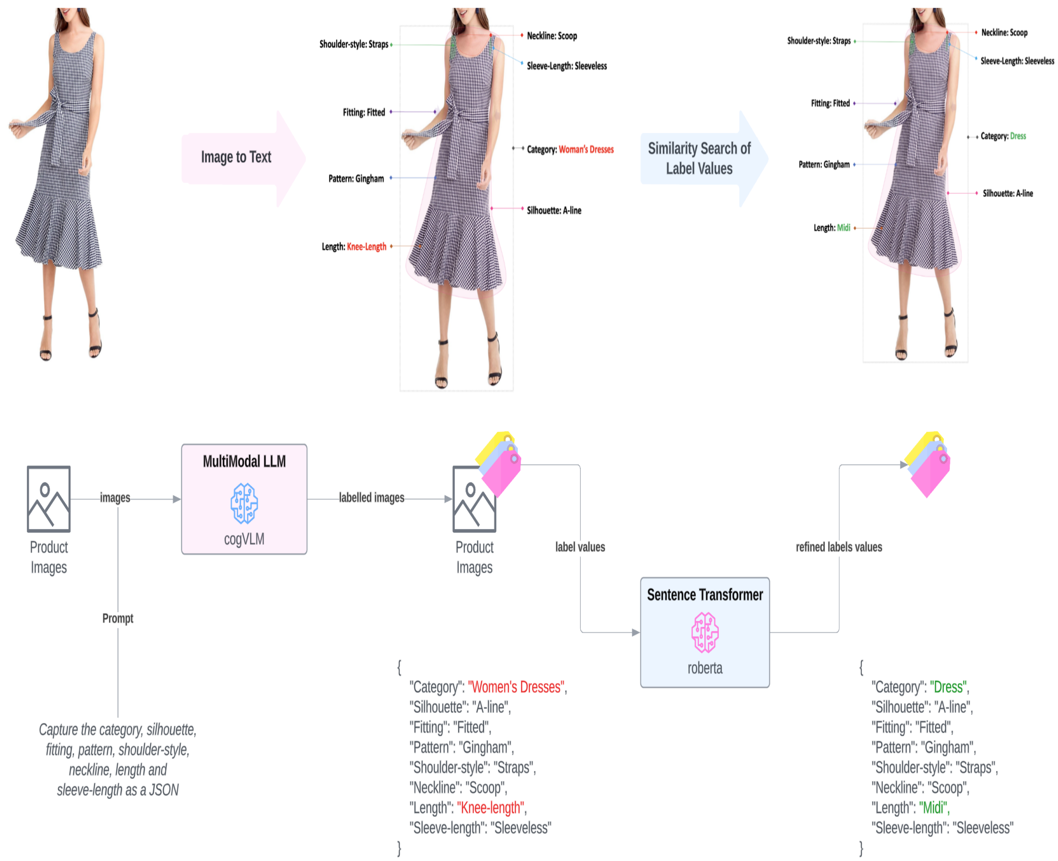
Observations: This approach has shown inconsistent results across images despite multiple iterations. The prompt design and possible value list (i.e., possible label values) only partially capture the nuances of the image attributes, leading to variability in model predictions.

Approach B: Conversational style prompt with possible value list



Description: This approach uses a conversational style prompt, where the model is queried about each image attribute individually, specifying the possible value list (i.e., possible label values). The prompt engages the model in a question-and-answer format to extract attributes incrementally. The study by Yusu Qian[6] shows that deceptive prompts exploit the sensitivity of Multimodel LMMs by introducing subtle linguistic cues that mislead the model. Therefore, we experimented with various prompt structures until we identified one that consistently yielded the most accurate and reliable results
Observations: While Approach B performs well for a majority of attributes, it did not achieve the same level of consistency and accuracy as the main approach(Approach C) for certain attributes. Some attributes exhibited lower performance or inconsistencies in extraction results compared to the main approach.

Approach C: Prompt with only attributes in combination with similarity search



Description: This approach uses a prompt that specifies only the attributes to extract from the image on **cogVLM**, followed by running a similarity search using **RoBERTa** against the possible value list. **Observations:** Contrary to our initial expectations, this method surpassed the performance of approaches A and C. Allowing the model the autonomy to select keywords that most accurately describe the image contributes significantly to a more effective determination of image attributes. This behavior suggests that the model's ability to independently identify and align attributes with their most fitting descriptors can lead to superior attribute extraction outcomes than more rigidly defined approaches.

6.3 Experimental results

These evaluation metrics are derived from comparing both Approach B and Approach C against manually labeled data, providing a comprehensive assessment of their performance in attribute classification. Overall, it appears that the categories of "Pattern" and "Sleeve-length" have the highest classification performance across all metrics, while "Neckline" shows significant room for improvement. The performance on "Fitting", "Length", "Shoulder-Style", and "Silhouette" varies, with each showing different levels of precision and recall, indicating that there may be specific challenges in these categories that could be addressed to improve the classification model.

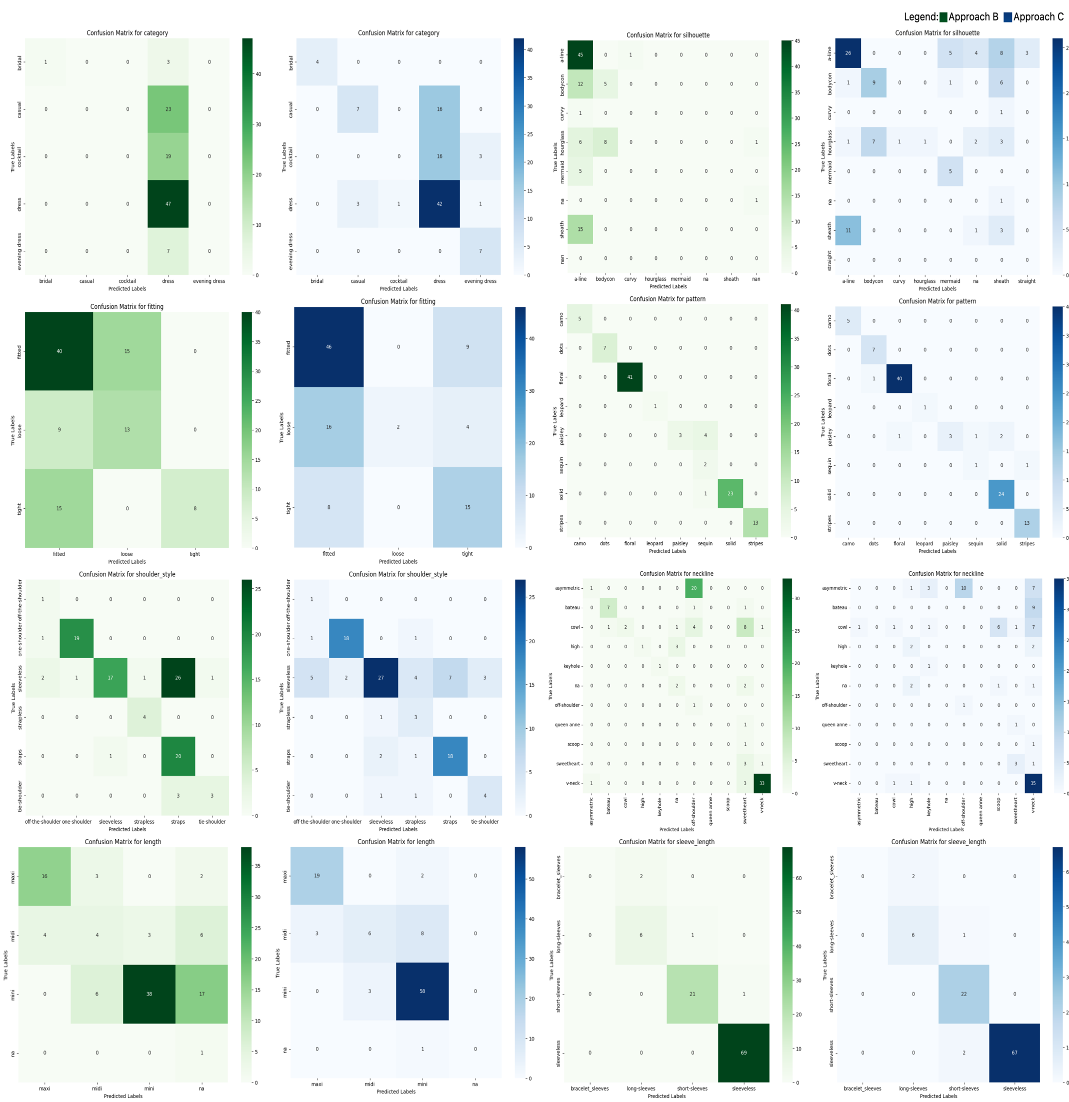
- **Comparative Analysis of Approaches B and C**
  Approach C performs better than approach B in several attributes such as category, silhouette, shoulder style, and length, indicating a more balanced and generally higher performance across various attributes. While excelling in pattern and sleeve length, Approach B shows weaknesses in other areas where Approach C takes the lead, especially in attributes like category and length, showing significant improvements.

Attributes/Evaluations	Precision	Recall	F1Score
Category	0.263	0.48	0.319
Silhouette	0.312	0.50	0.375
Fitting	0.676	0.61	0.603
Pattern	0.986	0.95	0.956
Shoulder-style	0.809	0.64	0.633
Neckline	0.773	0.51	0.521
Length	0.786	0.59	0.664
Sleeve-length	0.943	0.96	0.951

Table 1: Evaluation Metrics for Approach B

Attributes/Evaluations	Precision	Recall	F1 Score
Category	0.512	0.6	0.518
Silhouette	0.595	0.44	0.448
Fitting	0.705	0.63	0.577
Pattern	0.944	0.94	0.933
Shoulder-style	0.797	0.71	0.729
Neckline	0.331	0.43	0.325
Length	0.807	0.83	0.808
Sleeve-length	0.936	0.95	0.942

Table 2: Evaluation Metrics for Approach C



- **Comparative Analysis of Multi-modal LVMs**
  A comparative analysis of f1score and accuracy for manually labeled fashion images shows **cogVLM** performs the best, followed by **GPT4V**, **Qwen-VL** in a distant third, and LlaVa is significantly behind. The data suggests that while **cogVLM** and **GPT4V** are closely matched and highly effective, **Qwen-VL** and **LlaVa** offer decreased performance. **LlaVa** is notably less effective in the evaluated tasks.

Models	**cogVLM**	**GPT4V**	**LlaVa**	**Qwen-vl**
Avg. F1 Score	66.000	65.654	30.266	50.238
Avg. Accuracy	69.125	69.249	34.500	53.500

Table 3: Comparison of Multi-modal LVM Performances

6.4 Fine-tuning cogVLM

This chapter delves into fine-tuning the Swiss Army Transformer (SAT)[6] for image classification, enhanced with Low-Rank Adaptation [7], and optimized with DeepSpeed [8]. We focus on applying LoRA adjustments to the SAT model methodically, facilitating a targeted and efficient refinement of its attention mechanisms for improved image classification accuracy.

Fine-tuning with SAT and LoRA:

- **Model Adaptation:** Fine-tuning the Swiss Army Transformer (SAT) for a specific dataset incorporates LoRA with `-lora_rank 10`, enabling precise, low-rank matrix adjustments to the attention mechanisms. This process uses a smaller cosine decay learning rate (`-lr-decay-style cosine`) for nuanced weight adaptation.

- **Efficiency with DeepSpeed:** DeepSpeed optimizes training, making it feasible to handle the augmented computational demands of the LoRA-enhanced SAT model, allowing for efficient training on larger models or datasets.

- **Checkpointing:** Strategic checkpointing (`-save-interval 200`) is crucial for tracking LoRA's impact over the training phases, facilitating model recovery and periodic evaluations.

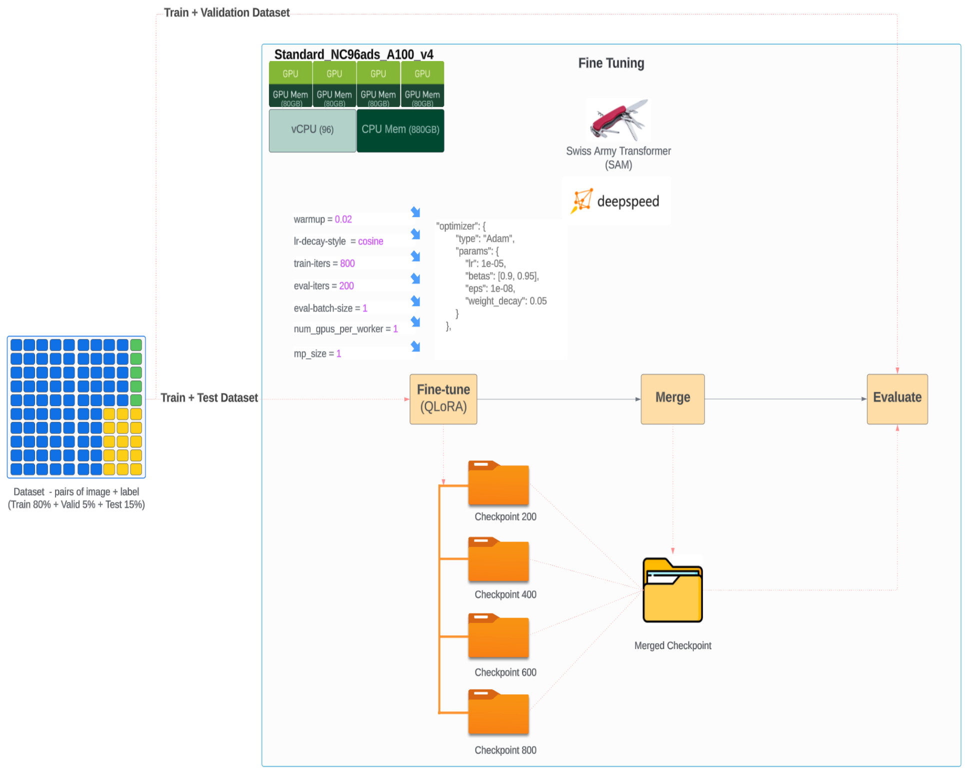
- **Model Merging and Evaluation:** Post-fine-tuning, we merge LoRA-modified components to ensure their uniform application across the model. We then evaluate this consolidated model to gauge performance improvements.

- **Parameter Configuration:** The process meticulously calibrates learning rates, weight decay, and epochs to optimize the LoRA-enhanced model's training. Specific adjustments ensure the model effectively learns from new data insights while preserving its foundational knowledge.

- **Targeted Training Parameters:** This approach utilizes a targeted approach with a warmup phase (`-warmup .02`) and specific epochs to fine-tune the SAT model with LoRA enhancements, aiming for optimal performance in image classification tasks.

  This fine-tuning approach involves careful parameter and configuration management, such as learning rate, weight decay, and epoch settings, which are pivotal for effectively training a deep learning model. By methodically adjusting these variables, we foster a structured pipeline capable of developing a formidable image classification system adept at precisely categorizing images with rich visual content.

[6] `https://github.com/THUDM/SwissArmyTransformer`



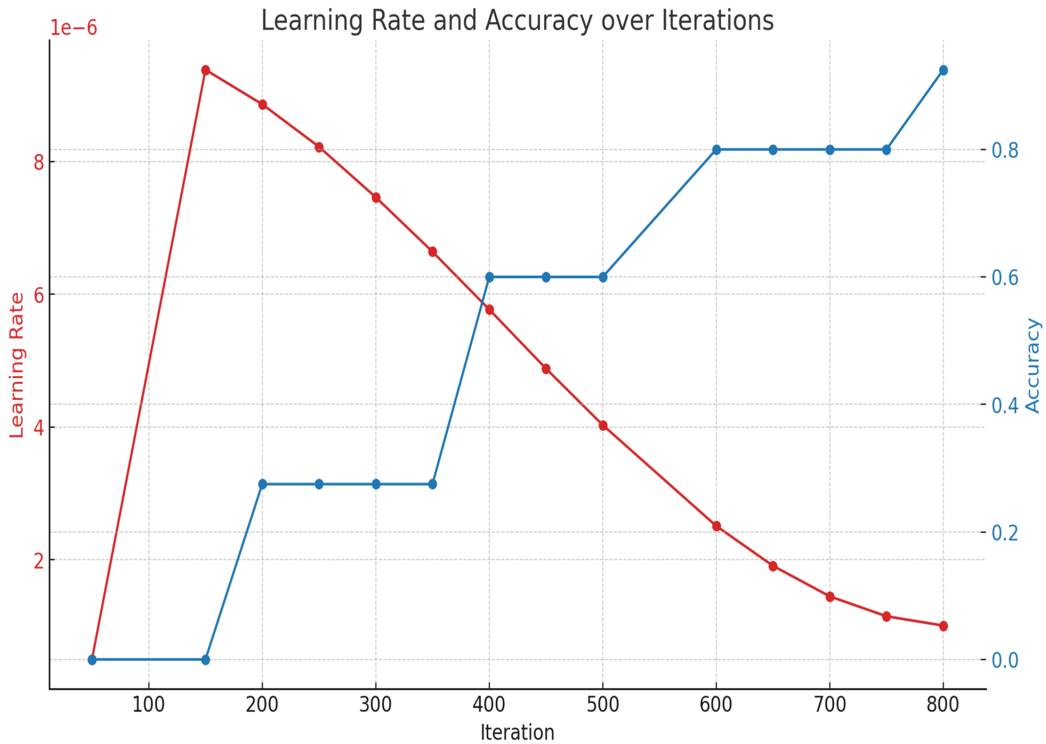
Fine-tuning Metrics:
The fine-tuning process has shown that the cogVLM adapts well over time, with a clear inverse relationship between learning rate and accuracy metrics. Notably, the accuracy improvement without case sensitivity is remarkable, highlighting the model's ability to learn invariant features. The total loss reduction signifies a successful fine-tuning phase, suggesting that the model has reached a point of convergence.

- The graph below presents the evolution of the learning rate and model accuracy over 800 iterations during the fine-tuning of a cogVLM using 2000 labeled images.

- The learning rate starts high and sharply decreases until around iteration 500, then levels off, indicating an initial aggressive learning strategy that becomes more conservative as the model optimizes. This is typical of adaptive learning rate strategies designed to converge efficiently.

- Simultaneously, the accuracy graph shows a stepwise increase, with notable improvements occurring at specific iterations. This suggests that certain updates during training significantly enhance model performance, possibly at points where the learning rate decreases.

- The final accuracy achieved is high, indicating a successful fine-tuning process in model performance on the given labeled image dataset.



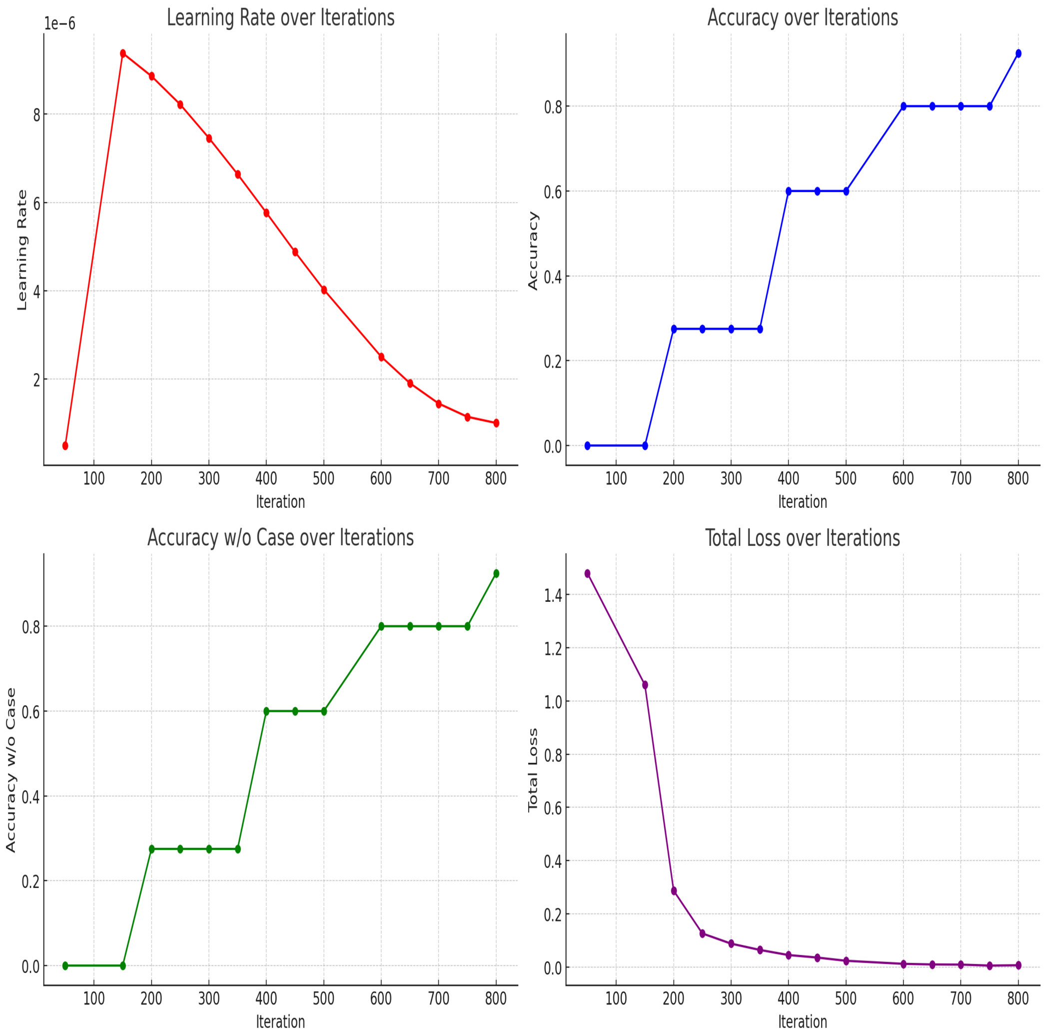
- The next set of graphs provides a more granular view of different aspects of the fine-tuning process.

- The first graph repeats the learning rate information, corroborating the strategy observed in the first set of graphs.

- The "Accuracy w/o Case" graph reveals a steady increase in model performance when case sensitivity is not a factor, suggesting that the model effectively learns general patterns and features from the invariant images to the text case.

- Lastly, the "Total Loss" graph shows a rapid decrease in loss during the initial iterations, which then plateaus. This indicates that the model quickly reaches a good level of performance and refines its understanding of the data incrementally.

- The plateauing of the loss suggests that the model may have reached its learning capacity, given the current architecture and dataset.



6.5 Quantitative Analysis

The fine-tuning process has particularly benefited attributes where the "pre-tuned" model exhibited notable deficiencies. It has significantly sharpened the accuracy of classifications within "Fitting," "Category," and "Neckline." This points to the model's improved comprehension of the intricacies inherent to these categories, leading to more precise predictions.

While "Silhouette" and "Shoulder Style" have seen moderate improvements with a decline in precision, additional fine-tuning or strategy modifications could optimize the precision-recall trade-off.

Overall, the model's robustness and reliability in "Length," "Sleeve Length," and "Pattern" classifications remain impressive, with fine-tuning solidifying their performance metrics. The amalgamation of insights from confusion matrices and precision, recall, and F1 scores showcases the substantial impact of fine-tuning in boosting model performance, particularly in domains where initial results indicated significant potential for enhancement. Ongoing refinements in the less-improved areas could pave the way for even greater accuracy and reliability in future predictions.

Average Performance Metrics: The overall improvements are quantifiable, with the average F1 score surging from 66.000 to 74.782 and the average accuracy escalating from 69.125 to 76.250 after fine-tuning. These improvements in aggregate metrics attest to the efficacy of fine-tuning in elevating the model's overall performance.

Models	Pre-trained cogVLM	Fine-tuned cogVLM
Average F1 Score	66.000	74.782
Average Accuracy	69.125	76.250

Table 4: Average Performance Metrics

- **Significant Improvements:**

  Fitting: The 'Fitting' attribute shows marked improvements, where a turnaround occurred from initially prevalent misclassifications, particularly within 'Tight' fittings. Post-fine-tuning, the model exhibited a notable uptick in "precision," "recall," and "F1 score," indicating enhanced precision and a higher hit rate in recognizing relevant instances.

  Category: The "Category" attribute witnessed a decrease in misclassification rates for "Casual" and "Cocktail" garments. The fine-tuned model improved precision and "F1 score," highlighting an enhanced ability to correctly identify these categories despite a marginal reduction in "recall."

  Neckline: Precision and F1 score for the "Neckline" attribute experienced a substantial improvement, signifying that fine-tuning markedly refined the model's accuracy and equilibrium between precision and recall.

- **Moderate Improvements:**

  Silhouette: Post-fine-tuning, the model's prowess in classifying "Silhouette" has advanced, significantly reducing "A-line" misclassifications. An increase in recall and F1 score indicates improved detection of true positives, although a dip in precision somewhat offsets this.

  Shoulder Style: Enhancements in the "Shoulder Style" classification are underscored by increased recall and F1 score, suggesting a heightened ability to flag relevant instances correctly—nonetheless, a marginal dip in precision points to a slight increase in false positives.

- **Minor or No Clear Improvement:**

  Length and Sleeve Length: These attributes displayed robust metrics even before fine-tuning. Post-fine-tuning, minor enhancements in precision, and "F1 score" for "Length" and "Sleeve Length" affirm that the fine-tuning fine-tuned where improvements were less pronounced.

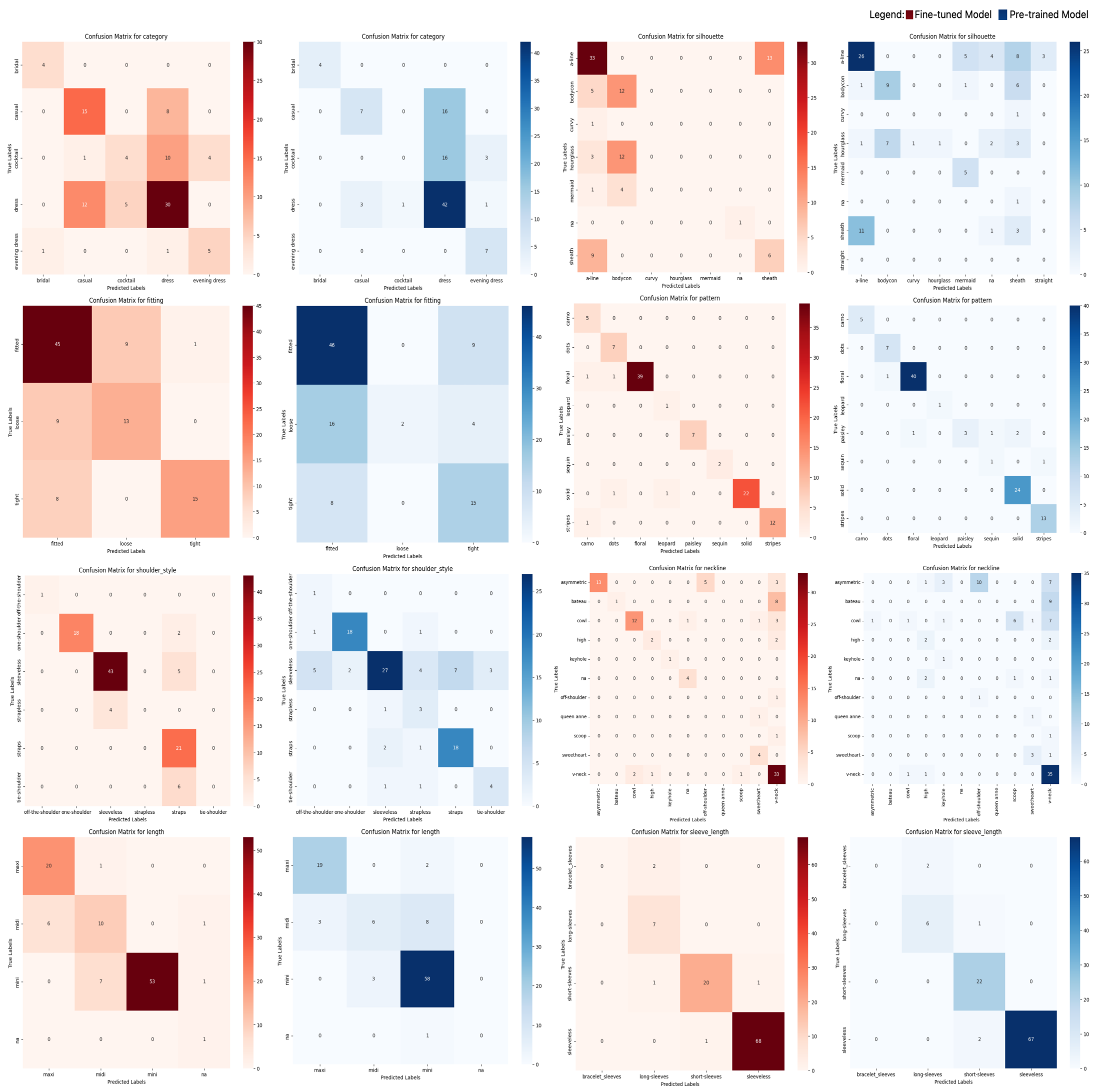
  Pattern: The "Pattern" attribute was already a stronghold of the model, and "fine-tuning" only yielded marginal gains. This underscores the model's pre-existing competence in pattern recognition, which is fine-tuning incrementally enhanced.

Attributes/Evaluations	Precision	Recall	F1 Score
Category	0.512	0.6	0.518
Silhouette	0.595	0.44	0.448
Fitting	0.704	0.63	0.576
Pattern	0.943	0.94	0.932
Shoulder-style	0.796	0.71	0.729
Neckline	0.330	0.43	0.324
Length	0.807	0.83	0.808
Sleeve-length	0.936	0.95	0.941

Table 5: Pre-trained cogVLM

Attributes/Evaluations	Precision	Recall	F1 Score
category	0.566	0.58	0.562
silhouette	0.422	0.52	0.463
Fitting	0.744	0.73	0.73
pattern	0.965	0.95	0.953
shoulder-style	0.778	0.83	0.794
neckline	0.780	0.7	0.688
length	0.869	0.84	0.848
sleeve-length	0.938	0.95	0.942

Table 6: Fine-tuned cogVLM



7 Qualitative Analysis

In this report, we conduct a qualitative analysis by manually comparing the results of our pre-trained and fine-tuned machine learning models against ground truth, represented by manually labeled images. This approach allows us to gain valuable insights into the performance and effectiveness of our models across different attributes.

- **Complete Alignment with Ground Truth**: In the first set, the fine-tuned cogVLM matches the ground truth perfectly across all attributes. This indicates that fine-tuning has likely adapted the model to this specific domain (dress attributes), improving accuracy. The pre-trained cogVLM has also performed well. However, the fine-tuning process has refined the model's predictions to align with the ground truth, showing that the fine-tuning process can effectively leverage additional domain-specific data to enhance the model's predictive capabilities.

- **Enhanced Accuracy through Fine-Tuning**: In the second set, the fine-tuned cogVLM improves over the pre-trained model in several areas. For instance, in the first example, the pre-trained model incorrectly identifies the dress fitting, whereas the fine-tuned model aligns with the ground truth. This indicates that fine-tuning has helped the model correct specific misclassifications.

- **Variable Model Performance Across Attributes**: In the third set, we observe cases where both models perform inconsistently. For instance, the pre-trained model might accurately identify one attribute but fail on another, and the fine-tuned model shows a similar pattern. This inconsistency could be due to various factors, including ambiguous image features, training data outliers, or model architecture limitations. This indicates that challenges still need to be addressed and that continuous improvement and adjustment of the models are necessary.

The fine-tuned cogVLM generally improves over the pre-trained model, aligning more closely with the ground truth data. However, there are instances where it fails or does not improve upon the pre-trained model's predictions, highlighting the complexity of the task and the need for ongoing model refinement.

Product Image	**Ground Truth**	**Pre-Trained cogVLM**	**Fine-Tuned cogVLM**

	{
	"category": "dress",
	"silhouette": "a-line",
	"fitting": "fitted",
	"pattern": "floral",
	"shoulder-style": "sleeveless",
	"neckline": "v-neck",
	"length": "mini",
	"sleeve-length": "short-sleeves"
	}
		{
		"category": "dress",
		"silhouette": "a-line",
		"fitting": "fitted",
		"pattern": "floral",
		"shoulder-style": "sleeveless",
		"neckline": "v-neck",
		"length": "mini",
		"sleeve-length": "short-sleeves"
		}
			{
			"category": "dress",
			"silhouette": "a-line",
			"fitting": "fitted",
			"pattern": "floral",
			"shoulder-style": "sleeveless",
			"neckline": "v-neck",
			"length": "mini",
			"sleeve-length": "short-sleeves"
			}

	{
	"category": "dress",
	"silhouette": "a-line",
	"fitting": "fitted",
	"pattern": "floral",
	"shoulder-style": "sleeveless",
	"neckline": "v-neck",
	"length": "mini",
	"sleeve-length": "short-sleeves"
	}
		{
		"category": "dress",
		"silhouette": "a-line",
		"fitting": "fitted",
		"pattern": "floral",
		"shoulder-style": "sleeveless",
		"neckline": "v-neck",
		"length": "mini",
		"sleeve-length": "short-sleeves"
		}
			{
			"category": "dress",
			"silhouette": "a-line",
			"fitting": "fitted",
			"pattern": "floral",
			"shoulder-style": "sleeveless",
			"neckline": "v-neck",
			"length": "mini",
			"sleeve-length": "short-sleeves"
			}

	{
	"category": "dress",
	"silhouette": "a-line",
	"fitting": "fitted",
	"pattern": "floral",
	"shoulder-style": "sleeveless",
	"neckline": "v-neck",
	"length": "mini",
	"sleeve-length": "short-sleeves"
	}
		{
		"category": "dress",
		"silhouette": a-line",
		"fitting": "fitted",
		"pattern": "floral",
		"shoulder-style": "sleeveless",
		"neckline": "v-neck",
		"length": "mini",
		"sleeve-length": "short-sleeves"
		}
			{
			"category": "dress",
			"silhouette": a-line",
			"fitting": "fitted",
			"pattern": "floral",
			"shoulder-style": "sleeveless",
			"neckline": "v-neck",
			"length": "mini",
			"sleeve-length": "short-sleeves"
			}

Complete Alignment with Ground Truth

Product Image	Ground Truth	Pre-Trained cogVLM	Fine-Tuned cogVLM

	{
	"category": "evening dress",
	"silhouette": "a-line",
	"fitting": "fitted",
	"pattern": "solid",
	"shoulder-style": "strapless",
	"neckline": "v-neck",
	"length": "maxi",
	"sleeve-length": "sleeveless"
	}
		{
		"category": "evening dress",
		"silhouette": "a-line",
		"fitting": "tight",
		"pattern": "solid",
		"shoulder-style": "strapless",
		"neckline": "v-neck",
		"length": "maxi",
		"sleeve-length": "sleeveless"
		}
			{
			"category": "evening dress",
			"silhouette": "a-line",
			"fitting": "fitted",
			"pattern": "solid",
			"shoulder-style": "strapless",
			"neckline": "v-neck",
			"length": "maxi",
			"sleeve-length": "sleeveless"
			}

	{
	"category": "dress",
	"silhouette": "bodycon",
	"fitting": "tight",
	"pattern": "sequin",
	"shoulder-style": "sleeveless",
	"neckline": "cowl",
	"length": "mini",
	"sleeve-length": "long-sleeves"
	}
		{
		"category": "dress",
		"silhouette": "bodycon",
		"fitting": "tight",
		"pattern": "sequin",
		"shoulder-style": "strapless",
		"neckline": "v-neck",
		"length": "midi",
		"sleeve-length": "long-sleeves"
		}
			{
			"category": "dress",
			"silhouette": "bodycon",
			"fitting": "tight",
			"pattern": "sequin",
			"shoulder-style": "sleeveless",
			"neckline": "cowl",
			"length": "mini",
			"sleeve-length": "long-sleeves"
			}

	{
	"category": "dress",
	"silhouette": "sheath",
	"fitting": "loose",
	"pattern": "floral",
	"shoulder-style": "sleeveless",
	"neckline": "high",
	"length": "mini",
	"sleeve-length": "long-sleeves"
	}
		{
		"category": "dress",
		"silhouette": "a-line",
		"fitting": "fitted",
		"pattern": "floral",
		"shoulder-style": "sleeveless",
		"neckline": "high",
		"length": "mini",
		"sleeve-length": "long-sleeves"
		}
			{
			"category": "dress",
			"silhouette": "sheath",
			"fitting": "loose",
			"pattern": "floral",
			"shoulder-style": "sleeveless",
			"neckline": "high",
			"length": "mini",
			"sleeve-length": "long-sleeves"
			}

Enhanced Accuracy through Fine-Tuning

Product Image	Ground Truth	Pre-Trained cogVLM	Fine-Tuned cogVLM

	{
	"category": "evening dress",
	"silhouette": "a-line",
	"fitting": "fitted",
	"pattern": "floral",
	"shoulder-style": "sleeveless",
	"neckline": "bateau",
	"length": "maxi",
	"sleeve-length": "sleeveless"
	}
		{
		"category": "evening dress",
		"silhouette": "mermaid",
		"fitting": "fitted",
		"pattern": "floral",
		"shoulder-style": "sleeveless",
		"neckline": "v-neck",
		"length": "maxi",
		"sleeve-length": "sleeveless"
		}
			{
			"category": "dress",
			"silhouette": "a-line",
			"fitting": "fitted",
			"pattern": "floral",
			"shoulder-style": "sleeveless",
			"neckline": "v-neck",
			"length": "maxi",
			"sleeve-length": "sleeveless"
			}

	{
	"category": "evening-dress",
	"silhouette": "mermaid",
	"fitting": "tight",
	"pattern": "solid",
	"shoulder-style": "one-shoulder",
	"neckline": "asymmetric",
	"length": "maxi",
	"sleeve-length": "sleeveless"
	}
		{
		"category": "evening-dress",
		"silhouette": "mermaid",
		"fitting": "tight",
		"pattern": "solid",
		"shoulder-style": "one-shoulder",
		"neckline": "off-shoulder",
		"length": "maxi",
		"sleeve-length": "sleeveless"
		}
			{
			"category": "evening-dress",
			"silhouette": "bodycon",
			"fitting": "tight",
			"pattern": "solid",
			"shoulder-style": "one-shoulder",
			"neckline": "asymmetric",
			"length": "maxi",
			"sleeve-length": "sleeveless"
			}

	{
	"category": "cocktail",
	"silhouette": "hourglass",
	"fitting": "tight",
	"pattern": "floral",
	"shoulder-style": "one-shoulder",
	"neckline": "asymmetric",
	"length": "mini",
	"sleeve-length": "sleeveless"
	}
		{
		"category": "dresses",
		"silhouette": "bodycon",
		"fitting": "fitted",
		"pattern": "floral",
		"shoulder-style": "one-shoulder",
		"neckline": "off-shoulder",
		"length": "mini",
		"sleeve-length": "sleeveless"
		}
			{
			"category": "dress",
			"silhouette": "bodycon",
			"fitting": "fitted",
			"pattern": "floral",
			"shoulder_style": "one-shoulder",
			"neckline": "asymmetric",
			"length": "mini",
			"sleeve-length": "sleeveless"
			}

Variable Model Performance Across Attributes

8 Conclusion

In this project, we delved into three methodologies using cogVLM for fashion attribute extraction from images, each demonstrating varied effectiveness. Approach A faced challenges with stability, particularly in differentiating between attribute names and values, underscoring the importance of precise attribute categorization. Contrarily, Approaches B and C exhibited robust performance, with C edging out B in precision, recall, and F1 scores, highlighting the potential for model optimization. Our comparative analysis further established cogVLM's superiority over other models like GPT4-V, LLaVA 1.5, and Qwen-VL, based on a dataset of 100 manually labeled dresses, affirming its fashion attribute extraction prowess.

Quantitative and qualitative evaluations showcased the fine-tuned cogVLM's enhanced performance, achieving over a 10% improvement in accuracy and a closer alignment with ground truth compared to its pre-trained version. This enhancement emphasizes the critical role of fine-tuning in refining model precision for domain-specific accuracy. Nonetheless, it also illuminated the limitations of fine-tuning; despite its considerable benefits, the model encountered occasional inaccuracies and displayed variable performance across different attributes, indicating avenues for further model refinement.

The findings reveal promising prospects for amplifying model performance through additional fine-tuning, suggesting potential improvements in Approaches A and B with more extensive training datasets. Advocating for the continued development of cogVLM, our research highlights the necessity for advanced fine-tuning techniques and the incorporation of diverse datasets to navigate the complexities inherent in fashion attribute extraction more adeptly.

Our investigation paves the way for future advancements in AI-powered fashion attribute extraction. By spotlighting the immediate achievements and outlining substantial opportunities for progress with ongoing model enhancements and dataset expansion, this research provides a clear trajectory for elevating cogVLM's capability in attribute extraction tasks, promising increasingly superior outcomes.

9 Team contributions

Nishant focused on constructing the model inference and executing fine-tuning process, while SiYi worked on data labelling and using RoBERTa for similarity task. Both worked on the final report together.

References

[1] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. https://arxiv.org/abs/2311.03079, 2023.

[2] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. https://arxiv.org/abs/2312.08914, 2023.

[3] Jiawei Zhang Tianyu Pang Chao Du Yi Ren Bo LiMin Lin. Benchmarking large multimodal models against common corruptions. https://arxiv.org/abs/2401.11943, 2024.

[4] Sukesh Perla and Johannes Kolbe. Image to text. https://huggingface.co/tasks/image-to-text.

[5] Menglin Jia et al. Fashionpedia. https://fashionpedia.github.io/home/, 2020.

[6] Yusu Qian et al. How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. https://arxiv.org/html/2402.13220v1, 2024.

[7] Edward Hu et al. Lora: Low-rank adaptation of large language models. arXiv:2106.09685, June 2021. https://arxiv.org/pdf/2106.09685.pdf.

[8] Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, Connor Holmes, Cheng Li, and Yuxiong He. Deepspeed data efficiency: Improving deep learning model quality and training efficiency via efficient data sampling and routing. arXiv:2212.03597v3 [cs.LG], January 2024. https://arxiv.org/abs/2212.03597v3.

A Appendix

We have included some of the materials we created for defining fashion dress attributes, understanding fashion dress attributes, construction of image and label pairs, the model configuration for the fine tuned model saved locally and Gradio based application.

Labeling for Model Training

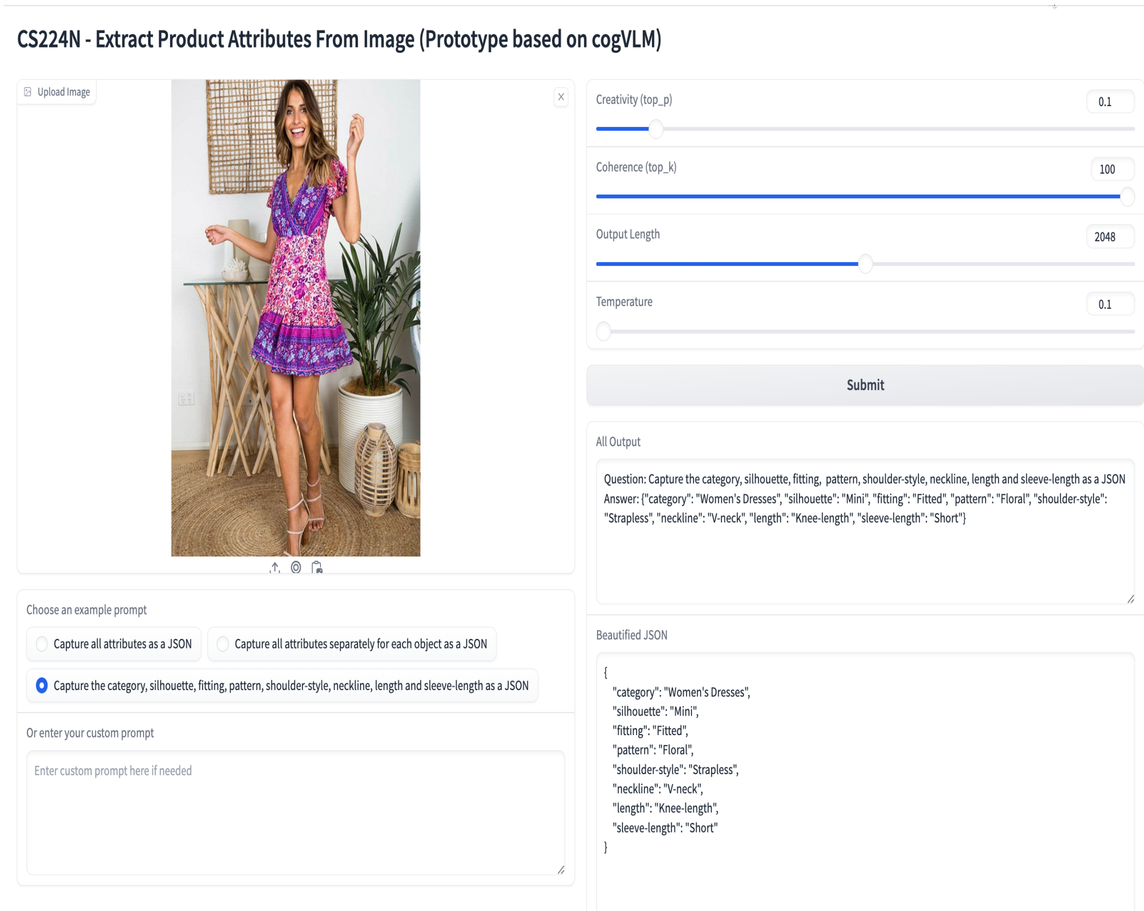
```
{
  "Question": "Capture the category, silhouette, fitting, pattern, shoulder-style, neckline, length and sleeve-length as a JSON",
  "Answer": {
    "category": "casual",
    "silhouette": "sheath",
    "fitting": "loose",
    "pattern": "gingham",
    "shoulder_style": "sleeveless",
    "neckline": "collar",
    "length": "mini",
    "sleeve-length": "short-sleeves"
  }
}
```

```
{
  "Question": "Capture the category, silhouette, fitting, pattern, shoulder-style, neckline, length and sleeve-length as a JSON",
  "Answer": {
    "category": "dress",
    "silhouette": "bodycon",
    "fitting": "tight",
    "pattern": "leopard",
    "shoulder_style": "sleeveless",
    "neckline": "plunging-v",
    "length": "maxi",
    "sleeve-length": "sleeveless"
  }
}
```

```
{
  "Question": "Capture the category, silhouette, fitting, pattern, shoulder-style, neckline, length and sleeve-length as a JSON",
  "Answer": {
    "category": "casual",
    "silhouette": "a-line",
    "fitting": "fitted",
    "pattern": "paisley",
    "shoulder_style": "sleeveless",
    "neckline": "keyhole",
    "length": "maxi",
    "sleeve-length": "sleeveless"
  }
}
```

Swiss Army Transformer - Fine Tuned Model Configuration:

```
{
  "model_class": "FineTuneTrainCogVLMModel",
  "tokenizer_type": "vicuna-7b-v1.5",
  "num_layers": 32,
  "hidden_size": 4096,
  "num_attention_heads": 32,
  "vocab_size": 32000,
  "layernorm_order": "pre",
  "model_parallel_size": 1,
  "max_sequence_length": 4096,
  "use_bias": false,
  "inner_hidden_size": 11008,
  "pre_seq_len": 8,
  "lora_rank": 10,
  "use_ptuning": false,
  "use_lora": true,
  "use_qlora": false,
  "layer_range": null,
  "image_length": 1225,
  "eva_args": {
    "model_class": "EVA2CLIPModel",
    "num_layers": 63,
    "hidden_size": 1792,
    "num_attention_heads": 16,
    "vocab_size": 1,
    "layernorm_order": "post",
    "model_parallel_size": 1,
    "max_sequence_length": 1226,
    "inner_hidden_size": 15360,
    "use_final_layernorm": false,
    "layernorm_epsilon": 1e-06,
    "row_parallel_linear_final_bias": false,
    "image_size": [
      490,
      490
    ],
    "pre_len": 1,
    "post_len": 0,
    "in_channels": 3,
    "patch_size": 14
  },
  "bos_token_id": 1,
  "eos_token_id": 2,
  "pad_token_id": 0,
  "image_size": 490
}
```


CS224N - Extract Product Attributes From Image (Prototype based on cogVLM)
Upload Image
Creativity (top_p)
0.1
Coherence (top_k)
100
Output Length
2048
Temperature
0.1
Submit
All Output
Question: Capture the category, silhouette, fitting, pattern, shoulder-style, neckline, length and sleeve-length as a JSON
Answer: {"category": "Women's Dresses", "silhouette": "Mini", "fitting": "Fitted", "pattern": "Floral", "shoulder-style": "Strapless", "neckline": "V-neck", "length": "Knee-length", "sleeve-length": "Short"}
Choose an example prompt
Capture all attributes as a JSON
Capture all attributes separately for each object as a JSON
Capture the category, silhouette, fitting, pattern, shoulder-style, neckline, length and sleeve-length as a JSON
Or enter your custom prompt
Enter custom prompt here if needed
Beautified JSON
{
"category": "Women's Dresses",
"silhouette": "Mini",
"fitting": "Fitted",
"pattern": "Floral",
"shoulder-style": "Strapless",
"neckline": "V-neck",
"length": "Knee-length",
"sleeve-length": "Short"
}