# An analysis of vision-language models for fabric retrieval

Francesco Giuliari[*,†], Asif Khan Pattan[†], Mohamed Lamine Mekhalfi[†] and Fabio Poiesi[†]

*Fondazione Bruno Kessler, Via Sommarive 18 - Povo, 38123 Trento, Italy*

## Abstract

Effective cross-modal retrieval is essential for applications like information retrieval and recommendation systems, particularly in specialized domains such as manufacturing, where product information often consists of visual samples paired with a textual description. This paper investigates the use of Vision Language Models(VLMs) for zero-shot text-to-image retrieval on fabric samples. We address the lack of publicly available datasets by introducing an automated annotation pipeline that uses Multimodal Large Language Models (MLLMs) to generate two types of textual descriptions: freeform natural language and structured attribute-based descriptions. We produce these descriptions to evaluate retrieval performance across three Vision-Language Models: CLIP, LAION-CLIP, and Meta's Perception Encoder. Our experiments demonstrate that structured, attribute-rich descriptions significantly enhance retrieval accuracy, particularly for visually complex fabric classes, with the Perception Encoder outperforming other models due to its robust feature alignment capabilities. However, zero-shot retrieval remains challenging in this fine-grained domain, underscoring the need for domain-adapted approaches. Our findings highlight the importance of combining technical textual descriptions with advanced VLMs to optimize cross-modal retrieval in industrial applications.

## Keywords

Cross-modal retrieval, text-to-image retrieval, fabric industry, vision-language models, automated annotation

## 1. Introduction

The retrieval of relevant content from databases is a fundamental task crucial for applications like information retrieval and recommendation systems. Effective retrieval enables efficient access to the vast amounts of daily generated data, allowing users to quickly find matching information or items [1].

Recent advancements in deep learning, particularly the development of aligned language and visual representations through contrastive pretraining [2], have significantly improved cross-modal matching between images and textual descriptions [3]. This progress enables efficient retrieval of visual information using text queries and vice versa, opening new possibilities for applications in various domains. In the manufacturing industry, where vast amounts of data are available in the form of production sample images, such as those featured in online stores, this capability allows for seamless retrieval of visual content based on textual descriptions, streamlining search and analysis processes.

Building on our work within the PNRR-funded "Intrecci Digitali" project, this paper focuses on zero-shot text-to-image retrieval, investigating how manipulating textual queries can enhance the retrieval accuracy of fabric samples. To conduct our study on text-to-image retrieval in the fabric domain, we identified a lack of suitable publicly available datasets. To address this, we utilized an existing fabric image dataset and generated corresponding textual prompts to enable our experiments. Specifically, we explore the use of MLLMs for automatically generating detailed image descriptions, allowing for a more robust evaluation of retrieval performance. We consider multiple types of textual prompts, including freeform and template-based descriptions incorporating fabric-specific attributes.

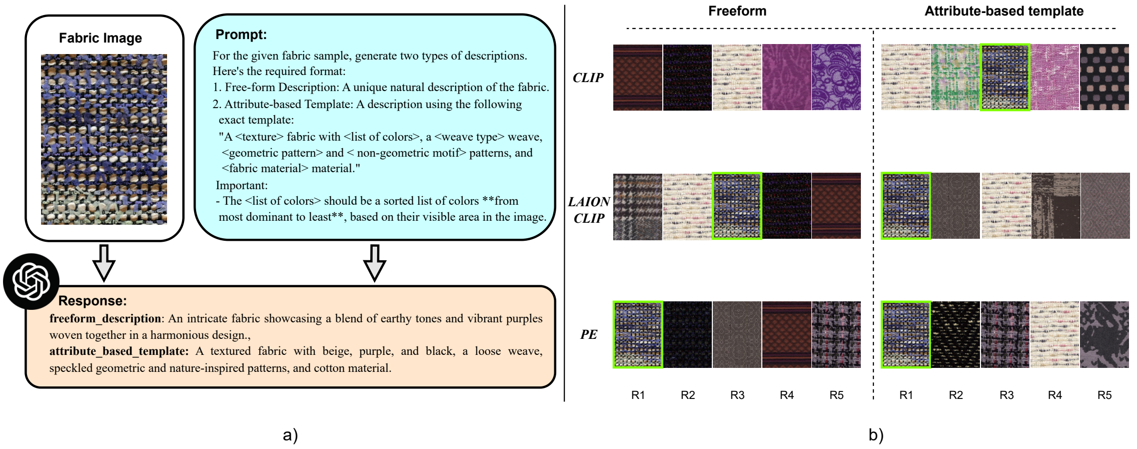
Our study highlights that to achieve the best retrieval accuracy, it is necessary to use both a technical attribute-based description and a powerful model that can effectively use such a description.

*Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy.*

[*] Corresponding author.

[†] All authors contributed equally.

fgiuliari@fbk.eu (F. Giuliari); apattan@fbk.eu (A. K. Pattan); mmekhalfi@fbk.eu (M. L. Mekhalfi); poiesi@fbk.eu (F. Poiesi)
0000-0003-2311-9517 (F. Giuliari); 0009-0008-9871-5931 (A. K. Pattan); 0000-0002-4295-0974 (M. L. Mekhalfi); 0000-0002-9769-1279 (F. Poiesi)



**Prompt:**

For the given fabric sample, generate two types of descriptions.
Here's the required format:
1. Free-form Description: A unique natural description of the fabric.
2. Attribute-based Template: A description using the following exact template:
"A <texture> fabric with <list of colors>, a <weave type> weave, <geometric pattern> and < non-geometric motif> patterns, and <fabric material> material."
Important:
- The <list of colors> should be a sorted list of colors **from most dominant to least**, based on their visible area in the image.

**Response:**

**freeform_description**: An intricate fabric showcasing a blend of earthy tones and vibrant purples woven together in a harmonious design.,
**attribute_based_template:** A textured fabric with beige, purple, and black, a loose weave, speckled geometric and nature-inspired patterns, and cotton material.

**Figure 1:** Visualization of our descriptions generation and retrieval process. (a) Shows the input fabric image, the prompt given to ChatGPT, and the resulting generated description. (b) Displays the Top-5 image retrieval results using two types of generated descriptions across three models: CLIP, LAION-CLIP, and the Perception Encoder (PE). R1, R2, etc., indicate the ranked order of retrieved images based on similarity scores.

## 2. Related Works

**Text-to-Image Retrieval** The development of VLMs like CLIP [2] has enabled effective cross-modal retrieval by learning to associate images and text within a shared embedding space. Trained on large-scale image-text datasets, these models align visual and textual information simultaneously, supporting tasks such as text-to-image retrieval. Earlier works [4, 5] in fabric image retrieval have primarily focused on image-to-image retrieval and are often limited to one or two specific attributes (e.g., texture, color). These approaches struggle with complex fabrics like printed textiles, where texture alone is insufficient to capture the full visual diversity. Recent advances in multimodal understanding-such as the pipeline proposed in [6]-demonstrate the effectiveness of combining MLLMs with VLMs for fine-grained visual reasoning. This integration enhances semantic alignment across modalities, which is crucial for cross-modal retrieval tasks requiring nuanced interpretation of visual attributes. Only a few recent studies, such as [7], have explored text-to-image fabric retrieval, proposing a training-based framework that leverages VLMs (e.g., CLIP [2]). In contrast, this work analyzes the zero-shot performance of VLMs on domain-specific cross-modal retrieval tasks such as fabric image retrieval, considering multiple visual attributes. Related analyses, such as [8], have shown that including expressive terms reflecting the 'tone' of a scene can enhance the alignment between text and image representations in natural images, suggesting the importance of descriptive richness in query formulation.

## 3. Analysis of cross-modal fabric retrieval

In cross-modal text-to-image retrieval, the goal is to identify an image in a database that corresponds to a given text description. Quantitatively evaluating this task requires high-quality data, as each image must be paired with a description that uniquely distinguishes it from all others. This challenge is particularly difficult for fabric images due to several factors: the lack of publicly available annotated datasets, the underrepresentation of fabric images in large-scale VLMs training data, and the inherent visual similarity among fabric samples, which makes them difficult to describe in a distinctive manner. To address the missing annotations in fabric image datasets, we introduce an automatic annotation pipeline that generates two distinct description types using LLMs: freeform natural language descriptions that describe the fabric holistically, and structured technical descriptions where predefined attribute templates are populated. In this study, we evaluate three VLMs using both description paradigms on fabric imagery, determining which combination of textual representation and embedding model proves most effective for this specific retrieval task.

**Dataset.** We use the Fabric-Image-Data (FID)[1][9], which comprises 12,181 wool fabric images categorized into four splits: lattice (3,128 images), pattern (768 images), solid (4,169 images), and stripe (4,116 images). Each image has a resolution of $420 \times 570$ pixels. While the original dataset is designed for fabric image classification, we adapt it for the task of text-to-image retrieval. To support this task, we generate two types of textual descriptions for each image. The detailed description generation process and the prompts used are described in the following section.

**Description Generation.** Descriptions are generated using ChatGPT-4o-mini based on the input prompt shown in Figure 1a. For the 'Attribute-based template' description, the attributes are selected from ChatGPT-4o-mini's response to a separate, image-independent prompt: *"List the top 5 attributes to distinguish a fabric image."* On average, the model generates each response in approximately 2.6 seconds.

**Experimental Setup.** We evaluate the retrieval performance of the models using both types of descriptions. For each model, we first pre-compute and store the image embeddings for all images in a given split. Then, for each textual description, we extract its text embedding using the model's text encoder. We compute the cosine similarity between the description's embedding and every image embedding in the split, ranking the results in descending order based on similarity. We report the results in terms of Hit-Rate @ Rank $K$ ($H@K$), where the $H@K$ for each query is 1 if the image corresponding to the description is among the first $K$ images and 0 otherwise. The final score is computed by averaging $H@K$ across all descriptions in the split. We report the scores with ranks: 1,5,10, and 20.

**Compared Vision-Language Models.** We compare three text-image embedding models: the CLIP[2][2] model from OpenAI; LAION-CLIP[3], which is a version of the CLIP model trained on LAION2B data[10]; and Perception Encoder[4][11], a recent model from Meta for image and text alignment.

**Results Discussion.** We evaluate the retrieval performance of three pre-trained models (CLIP, LAION-CLIP, and Perception Encoder) using two types of automatically generated descriptions: free-form and attribute-based template. Attribute-based descriptions were chosen for their ability to encode structured, fine-grained visual cues-such as color, weave, and pattern-that are critical for distinguishing highly similar fabric images. As shown in Figure 2, these structured descriptions lead to consistently higher retrieval accuracy across all models, with the greatest improvements observed in visually complex fabric classes like "lattice" and "printed." Among the models, the Perception Encoder delivers the strongest performance, benefiting significantly from attribute-based inputs. This suggests that its strong performance is driven by both a large training corpus and the use of intermediate feature representations, which enhance visual-text alignment through richer embeddings.

## 4. Conclusions

In this study, we report our findings on using pretrained Vision Language Models for Text-to-Image cross-modal retrieval on data from the fabric industry. Given the lack of manually annotated data in this domain, we propose a framework for the automatic labeling of fabric samples via the use of Chat-GPT to evaluate the performance of Text-to-Image retrieval systems.

Our findings highlight two critical factors for optimizing retrieval accuracy: First, retrieval accuracy improves substantially when fabric descriptions are technical and structured, incorporating details such as color, weave type, and patterns, rather than freeform, as evidenced by consistent gains across all tested models. Second, model choice plays a decisive role: while CLIP, the de facto standard academic VLM, performs poorly on fabric data even with augmented training (e.g., Laion-CLIP), the Perception encoder proves more adept at extracting relevant features from structured textual inputs. Nonetheless,

[1] https://github.com/rhrobot/Fabric-Image-Data

[2] https://huggingface.co/openai/clip-vit-large-patch14

[3] https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

[4] https://huggingface.co/facebook/PE-Core-L14-336

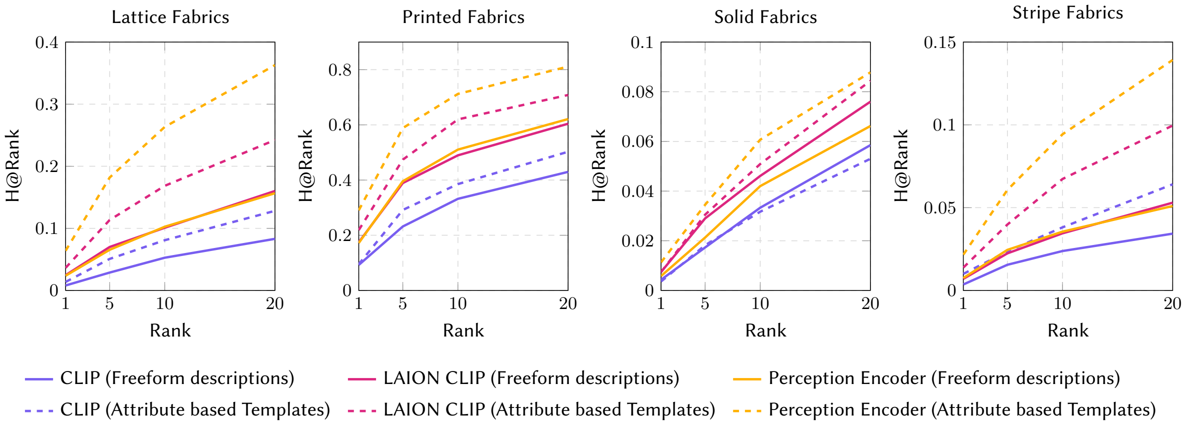


**Figure 2:** Retrieval Hit-rate analysis on the four fabric classes of FID. We compare three models: CLIP, LAION-CLIP, and Perception Encoder, using both Freeform descriptions and an Attribute-based template description.

absolute retrieval accuracy still is limited. Even with the most advanced model and descriptive input, zero-shot retrieval on fine-grained, domain-specific data such as fabrics remains a challenging task. In future works, we will continue to explore ways to improve performance in this particular domain.

## Acknowledgments

This work was carried out as part of the Intrecci Digitali project, developed within the research program PE00000004 'Made in Italy Circolare e Sostenibile – MICS', funded by the PNRR – Missione 4, Componente 2, Linea 1.3 – CUP D43C22003120001

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Deepseek in order to: Grammar and spelling check, Improve writing style, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, J.-R. Wen, Large language models for information retrieval: A survey, arXiv preprint arXiv:2308.07107 (2023).

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, et al., I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, 2021.

[3] L. V. B. Beltrán, J. C. Caicedo, N. Journet, M. Coustaty, F. Lecellier, A. Doucet, Deep multimodal learning for cross-modal retrieval: One model for all tasks, Pattern Recognition Letters 146 (2021).

[4] M. Wang, J. Wang, N. Zhang, J. Xiang, W. Gao, Fabric image retrieval based on decoupling of texture and color feature, Journal of Engineered Fibers and Fabrics 19 (2024) 15589250241246074.

[5] J. Xiang, N. Zhang, R. Pan, W. Gao, Efficient fine-texture image retrieval using deep multi-view hashing, Computers & Graphics 101 (2021) 93–105.

[6] M. Liu, S. Roy, W. Li, Z. Zhong, N. Sebe, E. Ricci, Democratizing fine-grained visual recognition with large language models, arXiv preprint arXiv:2401.13837 (2024).

[7] D. Suzuki, G. Irie, K. Aizawa, Text-to-image fashion retrieval with fabric textures, in: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, 2023, pp. 525–529.

[8] M. Sultan, L. Jacobs, A. Stylianou, R. Pless, Exploring clip for real world, text-based image retrieval, in: 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, 2023, pp. 1–6.

[9] R. Liu, Z. Yu, Q. Fan, Q. Sun, Z. Jiang, The improved method in fabric image classification using convolutional neural network, Multimedia Tools and Applications 83 (2024) 6909–6924.

[10] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, et al., LAION-5b: An open large-scale dataset for training next generation image-text models, in: 36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.

[11] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed, et al., Perception encoder: The best visual embeddings are not at the output of the network, arXiv preprint arXiv:2504.13181 (2025).