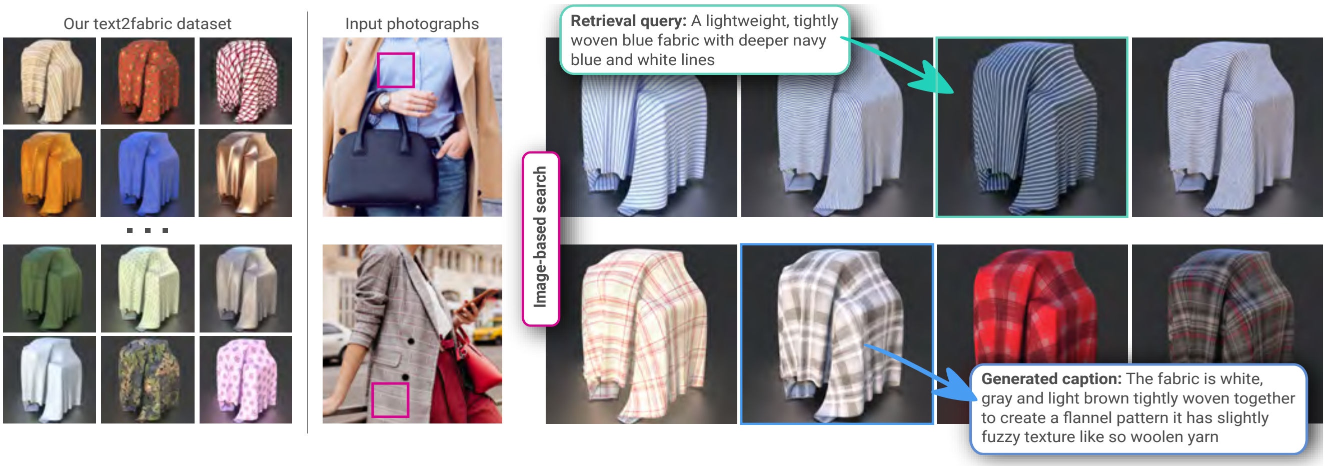
The Visual Language of Fabrics

VALENTIN DESCHAINTRE*, Adobe Research, UK
JULIA GUERRERO-VIU*, Universidad de Zaragoza - I3A, Spain
DIEGO GUTIERREZ, Universidad de Zaragoza - I3A, Spain
TAMY BOUBEKEUR, Adobe Research, France
BELEN MASIA, Universidad de Zaragoza - I3A, Spain



Retrieval query: A lightweight, tightly woven blue fabric with deeper navy blue and white lines

Generated caption: The fabric is white, gray and light brown tightly woven together to create a flannel pattern it has slightly fuzzy texture like so woolen yarn

Fig. 1. Our text2fabric dataset links high-quality renderings of a large variety of fabric materials to natural language descriptions of their appearance. We conduct a thorough analysis of this dataset, and leverage it to fine-tune large-scale vision-language models for a variety of tasks. We show here examples of such tasks: (i) image-based search, even using real photographs as input, yields relevant results from our dataset (the magenta square in the photographs marks the input crop, and the corresponding search results can be found in each row); (ii) text-based queries (green) result in successful fine-grained retrieval within the dataset; and, (iii) given an input image, we can generate detailed and rich descriptions of appearance (blue). Our work not only derives interesting insights regarding how people describe (fabric) appearance, but also demonstrates that a relatively small amount of high-quality data enables successful application of large vision-language models to specialized domains.

We introduce text2fabric, a novel dataset that links free-text descriptions to various fabric materials. The dataset comprises 15,000 natural language descriptions associated to 3,000 corresponding images of fabric materials. Traditionally, material descriptions come in the form of tags/keywords, which limits their expressivity, induces pre-existing knowledge of the appropriate vocabulary, and ultimately leads to a chopped description system. Therefore, we study the use of free-text as a more appropriate way to describe material appearance, taking the use case of fabrics as a common item that non-experts may often deal with. Based on the analysis of the dataset, we identify a compact lexicon, set of attributes and key structure that emerge from the descriptions. This allows us to accurately understand how people describe fabrics and draw directions for generalization to other types of materials. We also show that our dataset enables specializing large vision-language models such as CLIP, creating a meaningful latent space for fabric appearance, and significantly improving applications such as fine-grained material retrieval and automatic captioning.

CCS Concepts: • **Computing methodologies → Appearance and texture representations**.

Additional Key Words and Phrases: material appearance, perception, descriptions

1 INTRODUCTION

The recent quality surge in multimodal natural language processing (NLP) and vision-language models has enabled new interaction possibilities between images and text [Chen et al. 2022; Li et al. 2022; Poole et al. 2022; Radford et al. 2021; Ramesh et al. 2022; Saharia et al. 2022]. However, the underlying text-to-image models are trained on hundreds of millions of data points collected online, with a bias towards natural images and pictures typically found on the internet, and with general descriptions that do not capture the fine details and rich subtleties that domain-specific applications may require.

In this paper we explore the use of natural language to convey *fine-grained* material appearance. We pose three open key questions: (i) Is there an underlying common lexicon and structure when people describe material appearance using natural language? (ii) Can we communicate material appearance precisely enough with natural language only? (iii) Are language concepts relevant to material appearance well understood by large foundational models [Li et al. 2022; Radford et al. 2021]?

Although these questions are relevant for any type of vision-language model dealing with material appearance, to make the task

[*] Joint first authors

Authors' addresses: Valentin Deschaintre*, Adobe Research, UK, deschain@adobe.com; Julia Guerrero-Viu*, Universidad de Zaragoza - I3A, Spain, juliagviu@unizar.es; Diego Gutierrez, Universidad de Zaragoza - I3A, Spain, diegog@unizar.es; Tamy Boubekeur, Adobe Research, France, boubek@adobe.com; Belen Masia, Universidad de Zaragoza - I3A, Spain, bmasia@unizar.es.

tractable we focus on the particular class of *fabrics*. We choose this class since fabrics exhibit a wide variety of looks, patterns and reflectivity properties at different scales, are familiar to everyone, and are ubiquitous, widely present in many daily scenarios.

We first build a large dataset, text2fabric, relating photorealistic renderings of digital fabrics covering a wide range of appearance to natural language descriptions provided through crowdsourcing. We then perform a thorough analysis of our data, aimed at improving our understanding of how fabric appearance is described, and find that: (i) there is a common lexicon (ca. 500 words are enough to cover 95% of the 15,461 valid descriptions gathered); (ii) common properties (attributes) of appearance emerge from the descriptions; (iii) users do follow a certain structure when describing fabric appearance; and (iv) despite the infinite description space provided by natural language, there is a high similarity between descriptions of the same fabric. All these findings suggest that we have a shared understanding of language as it relates to material (fabric) appearance, which is key to communicating appearance precisely.

In addition, we leverage two successful and widely used vision-language models —CLIP [Radford et al. 2021] and BLIP [Li et al. 2022]— and show how their performance improves significantly when fine-tuned on our dataset. Last, we demonstrate applications of our model for various tasks (see Figure 1), including fine-grained text-based retrieval, image-based search, and automatic description or captioning of fabrics, all of them robust in the presence of light and geometry variations.

In summary, we present the following contributions:

- A text2fabric dataset including 15,000+ descriptions associated with 3,000 different fabric materials. The dataset is further augmented with 42,000 additional images featuring different geometries and lighting.
- A general methodology to collect and analyze natural language data describing images of fabrics, which is applicable to other domains.
- The identification of a common lexicon, structure and curated set of attributes that are relevant when describing fabrics.
- Fine-tuned models demonstrating the benefit of our dataset on several tasks.

Our full text2fabric dataset, as well as the fine-tuned models, are made publicly available to facilitate future research[1].

2 RELATED WORK

Description of visual attributes. Describing the appearance of objects, scenes or situations through language is a common task for humans. It allows to transmit richer information than simpler labeling and categorization approaches. Descriptions are not only more natural, but they also allow to focus on key or unusual aspects, or to add comparisons or semantics [Farhadi et al. 2009]. This information would then be leveraged by means of natural language processing (NLP) to enable new, more user-friendly computer graphics and vision algorithms. While providing a complete, general method for *any* object is still a daunting task, several methods have been proposed for people [Bourdev et al. 2011], faces [Kumar et al. 2011], or scenes [Patterson and Hays 2012]. Other authors have focused on the particular problem of texture description. Bhushan et al. [1997] came up with a limited 98-word lexicon which could describe 82% of their experimental data, formed by textures; instead of using text descriptions, participants had to cluster words based on perceived similarity. Inspired by this work, Cipoi and colleagues [2014] introduced the Describable Textures Dataset (DTD), composed of more 5,000 images labeled with one or more adjectives in a simple lexicon of 47 texture terms. The work was later extended into the Describable Textures in Detail Dataset (DTD^2) [Wu et al. 2020], including natural language descriptions. Recently, Xu et al. [2022] presented Texture BERT, a learning-based architecture that minimizes distances between texture and text features, optimized for image retrieval. While the domain of texture descriptions is rich and varied, our notion of appearance goes beyond 2D RGB maps, including aspects like reflectance, glossiness, touch, use, weight, etc. Our methodology further has the potential to generalize to other material classes.

Perceptually-meaningful material spaces. The role of perception in computer graphics has been extensively researched [Bartz et al. 2008; Fleming et al. 2015; McNamara et al. 2011; Thompson et al. 2011]. In particular, finding perceptually-meaningful material spaces has many applications in graphics, including editing [Serrano et al. 2016; Shi et al. 2021], gamut mapping [Sun et al. 2017], image-space manipulations [Boyadzhiev et al. 2015; Delanoy et al. 2022; Khan et al. 2006] or material similarity [Lagunas et al. 2019].

Pellacini et al. [2000] derived a two-dimensional perceptually uniform space for gloss, correlated with the parameteres of the Ward BRDF model [Ward 1992]; the concept was later extended by Wills and colleagues [2009] to include different reflectance models. Focusing on the problem of optimal reflectance acquisition, Nielsen et al. presented a perceptual scaling and decomposition of BRDF data, which allowed to reduce PCA dimensionality; the authors further showed how the first few dimensions roughly correlate with the specular and diffuse components of appearance [Nielsen et al. 2015]. In computer graphics, the joint effect of geometry and illumination on appearance has also been thoroughly studied [Bousseau et al. 2011; Dror et al. 2001; Lagunas et al. 2021; Storrs et al. 2021; Vangorp et al. 2007]. Recently, Serrano and colleagues [2021] trained a deep learning architecture using over 40,000 combinations or shape, material and illumination, to predict perceptual attributes of materials that correlate with human judgements.

Distilling human-centered knowledge. Understanding how people perform certain tasks and interact with different concepts is an important aspect of many human-centered computer graphics applications. Gathering rich, annotated datasets allows to distill this knowledge and apply it to the design of intuitive interfaces and workflows, help the development of novel algorithms, and automate time-consuming tasks. For instance, Cole et al. [2008] and Gryaditskaya et al. [2019] gathered a dataset of sketches and carefully analysed the practices of artists in terms of line types and how they are used. Garces et al. [2014] provided a measure of style similarity for clip art by gathering and analyzing human responses in a dataset of a thousand elements. The *OpenSurfaces* dataset [Li and Snavely 2018] contains crowdsourced pairwise comparisons of

[1] Project website: https://valentin.deschaintre.fr/text2fabric

material properties, to improve the performance on difficult tasks such as intrinsic image decomposition. Jarabo and colleagues [2014] tackled the problem of navigating the four-dimensional structure of light fields to provide an intuitive interface for editing them. Last, data from over 800 participants exploring VR scenes has been used to devise novel compression or video synopsis algorithms [Sitzmann et al. 2018].

We frame our data gathering and analysis in a similar fashion to these works. Our goal is to distill important knowledge about how people describe fabrics, show applications like automatic captioning and retrieval, and suggest a generalization of our methodology to a larger set of material classes.

3 OUR TEXT2FABRIC DATASET

This section describes how we designed and created our large-scale text2fabric dataset. It consists of 45,000 rendered images depicting samples of 3,000 different fabrics, together with 15,461 associated descriptions in natural language. The full dataset, with a web-based browser to explore it, can be downloaded from the project website: https://valentin.deschaintre.fr/text2fabric.

3.1 Rendered Images

The 45,000 images of our dataset are generated using the Substance Stager renderer, and span 3,000 different fabric materials with a wide range of appearance. These fabric models can be found on the Substance 3D website[2], and consist of both procedural materials generated by artists, as well as high-quality scans. Procedural materials come in the form of directed acyclic graphs, made of nodes of three different types: generators, which typically define the global bidimensional structure of the material (e.g., tiles); filters, which alter their input (e.g., colorization); and data stores, which point to external resources (e.g., raster content). Once executed by a material graph engine, these procedural models provide the material channels in the form of 2D maps, at a chosen resolution. A few parameters of the nodes of a graph are exposed as hyperparameters, so that changing them yields meaningful variations of the so-defined material. In the case of fabrics, these hyperparameters are carefully chosen so that the bounded variation they produce maps realistically to patterns and textile types encountered in the fabric industry. In addition, each procedural material originally includes a title and some tags describing its main appearance (e.g., sportswear, upholstery, mesh). We choose not to rely on these, as they describe what the artist wanted to represent rather than how people perceive it, do not follow any particular convention, and may introduce bias in the descriptions. Nonetheless, this information may be used to complement our collected descriptions.

For our task, we first rendered images of all 3,000 different materials at 4K resolution on a *baseline* geometry, carefully chosen to faithfully convey the appearance of the fabric: it contains both draped and flat areas, covering a wide range of orientations. We then selected a *baseline* indoor illumination, featuring soft lighting through multiple windows. A representative sample of the resulting fabrics can be seen in Figure 2 (top row).

Additionally, to ensure robustness and help future applications (see Section 5), we rendered the same materials using four other geometries—a sphere and a plane, in both draped and non-draped versions—, and two other illuminations—*outdoor*, with direct outdoor illumination, and *studio*, with strong studio indoor lighting, yielding 42,000 more images. Figure 2 (bottom row) shows some examples.

Different from other existing general-purpose datasets like ImageNet or LAION[3], our fabrics dataset constitutes a quite specific subset of images. This is illustrated, e.g., by the GLCM entropy [Haralick et al. 1973], a measure of randomness of the images, as shown in Figure 3a; our data yields a narrower histogram (narrower range of entropy), compared to the same number of randomly selected images from LAION (other image statistics can be found in the supplemental material). The specific characteristics of our image data are relevant to its use in learning-based models, such as the ones employed in Section 5.

3.2 Natural Language Descriptions

In the garment manufacturing industry, the description of fabrics involves specialized concepts and words such as *permeability* (how much air or water it allows through), *absorbency* (the ability of a fabric to take in moisture), or *colorfastness* (the ability of a fabric to maintain its color and resist fading), to name a few [glo 2001]. This specialized vocabulary is different from the one used by digital artists and practitioners in general. We thus gather our own text data to describe fabrics.

We collected 15,461 valid descriptions of fabric appearance as free text using natural language, through a carefully controlled crowdsourcing framework (Section 3.2.1), followed by a process of data verification and auditing (Section 3.2.2). Finally, we post-processed the resulting data (Section 3.2.3) in preparation for the analysis described in Section 4.

3.2.1 Annotation Procedure and Participants. We conducted a crowd-sourced user study in which participants (which we term *describers*) had to provide free-text descriptions for our high-quality fabric renderings. Specifically, describers were shown one image at a time, along with three zoomed-in areas (highlighted in green in the top-left image of Figure 2), and were asked to describe the appearance of the material as precisely as possible using their own words in natural language. Describers were free to use one or several sentences for the descriptions (1-3 was recommended), and word count was limited to the range [20..100] words, to prevent excessively short or long descriptions. To keep the task tractable, we gathered descriptions for our *baseline* set of 3,000 images of different materials. This also encouraged describers to focus on the only changing aspect between images –the material–, familiarizing themselves with the geometry and illumination. This decision is further justified by the nature of the task, the goal of our study, and the ability of the human visual system to achieve perceptual stability and extract constant properties of materials from varying viewing conditions [Fleming 2017; Fleming et al. 2015; Tsuda et al. 2020]. Describers were required to do a minimum of ten descriptions, and we ensured that

[2] https://substance3d.adobe.com/assets/allassets?assetType=substanceMaterial&category=Fabric

[3] LAION [Schuhmann et al. 2021] is the large-scale dataset on which the vision-language model CLIP [Radford et al. 2021] is trained.

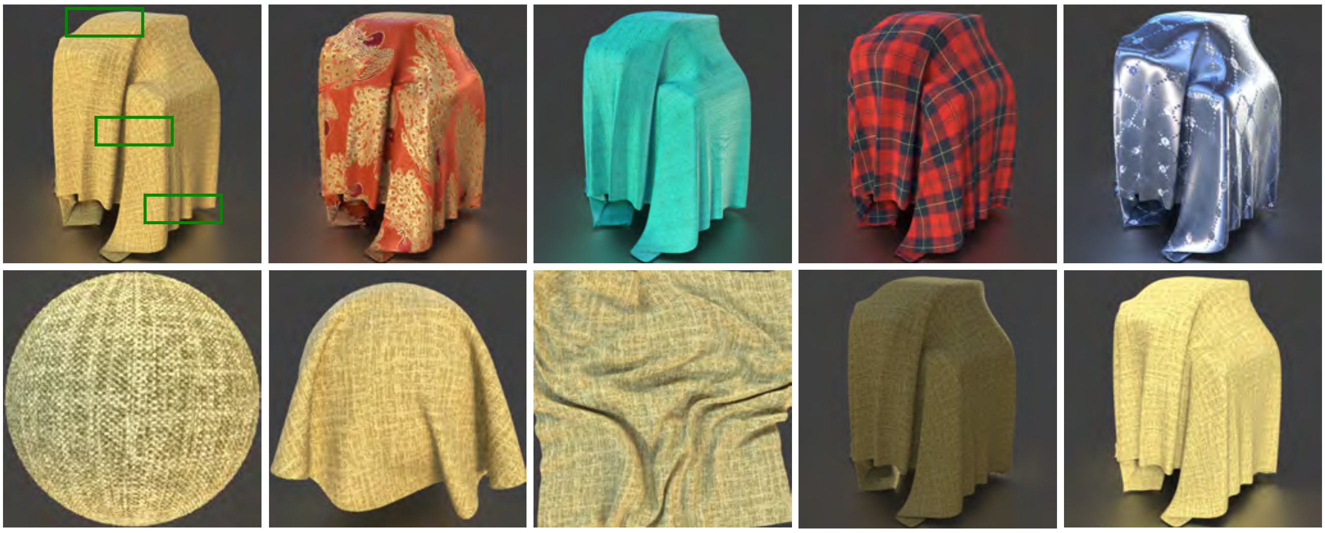


Fig. 2. Representative images from our text2fabric dataset. *Top row:* Five sample fabric materials, rendered with our *baseline* geometry and illumination. Highlighted areas in the first image mark zoomed-in regions shown to describers. *Bottom row:* Dataset images featuring three of our additional geometries and the two additional illuminations, all with the same fabric material (top row, leftmost material); from left to right: *sphere*, *sphere-draped* and *plane-draped* geometries, and *outdoor* and *studio* illuminations.

no describer contributed more than 9% of the whole text data. We also ensured that, for each image, we gathered at least five *valid* descriptions from different participants.

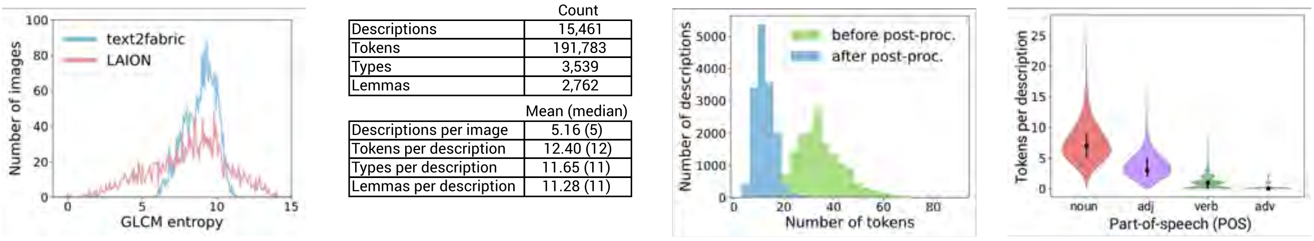
Given the specifics of the task, we required that the describers be native English speakers, had normal color vision, and were familiar with fashion or design. While we are aware that this may introduce some bias in the responses, it allows to gather a rich and accurate vocabulary. Prior to taking part in the study, participants underwent a short training and a qualification test. The training consisted of a set of instructions along with example descriptions gathered from a smaller pilot study. For the qualification test, each participant had to describe ten test fabric renderings; participants offering overly simple, poor descriptions, such as “this is a nice fabric”, were discarded. Approximately one in four did not pass this qualification test. After this process, a total of 122 describers (ages 18 through 65) went on to provide descriptions for our dataset: 45% identified as female, 12.3% as male, none as other gender identities, and 42.7% preferred not to reply.

3.2.2 Data Verification. We gathered a total of 19,167 descriptions from the 122 qualified describers. However, ensuring quality in free text description is a difficult task. We therefore established an additional continuous data verification protocol in which we manually audited descriptions. For each description, we first labeled them manually as one of four options: *accepted*, or rejected due to the description being *too generic*, being *wrong*, or using *poor grammar* to the point of hindering understandability. In addition, we also rated each description using a 5-point scale (1=totally unacceptable, 2=unacceptable, 3=acceptable, 4=very good, 5=excellent).

Manual auditing of the full set of almost 20,000 descriptions is an arduous task. However, we found that the quality of the descriptions was highly dependent on the describer, and data quality (as given by the ratings) was fairly uniform within a describer. Therefore, auditing a randomly-selected subset of the descriptions of a participant provided a good estimate for the rest of their descriptions; for example, for participants with a rejection rate > 35%, we rejected all their remaining, non-audited descriptions. More details of this process can be found in the supplemental material. The data gathering process was iterative, to ensure that we had at least five descriptions for each fabric. After this process we ended up with 15,461 valid descriptions and 3,706 invalid ones (a 19.3% rejection rate).

3.2.3 Post-processing. Following standard natural language processing techniques, we post-process our text data by removing non-alphabetic characters, applying a spell checker, and filtering stop words. Moreover, to carry out a proper analysis we extract tokens, types, and lemmas from the descriptions [Brezina 2018]. A *token* is each occurrence of a word in a text, while a *type* is each unique occurrence of a word in a text. A *lexeme* corresponds to the set of alternating forms from a common root word (such as “colors”, “colored” or “coloring”), while a *lemma* refers to the particular form chosen to represent a lexeme (such as “color” in our previous example). Further details, including the spell checker and the lemmatizer we use, can be found in the supplemental material.

The statistics of our textual data after this post-processing can be found in Figure 3b. When looking at values per description, we can see that the mean and median are close, indicating that the distributions are not too skewed; we show this distribution for the case of tokens in Figure 3c, both before and after post-processing. The table in Figure 3b (bottom) further shows that the difference between the number of tokens, types and lemmas per description is not large, suggesting that our descriptions are diverse in the sense that there are not many repeated words in them. Finally, we also perform part-of-speech tagging and classify tokens into nouns,



	Count
Descriptions	15,461
Tokens	191,783
Types	3,539
Lemmas	2,762

	Mean (median)
Descriptions per image	5.16 (5)
Tokens per description	12.40 (12)
Types per description	11.65 (11)
Lemmas per description	11.28 (11)

(a) (b) (c) (d)

Fig. 3. Statistics of our text2fabric dataset. (a) GLCM entropy distribution of our image data, and of a randomly-selected subset of the LAION dataset [Schuhmann et al. 2021] of the same size as ours, for comparison. (b) Statistics of our textual data. (c) Histogram of the length, in tokens, of our descriptions before and after post-processing the text. (d) Violin plots showing part-of-speech (POS) tagging for our descriptions (black lines show the IQR, and the median is indicated by a black dot).

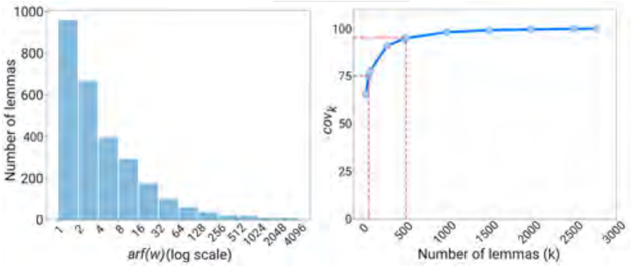


Fig. 4. Lexicon of fabric descriptions. *Left:* Histogram of average reduced frequency ($arf(w)$) of lemmas found in our descriptions; note that the x-axis is log-scale. *Right:* Mean coverage of descriptions (cov_k) for different levels of k: a description coverage of 75% is achieved with the most prominent 84 lemmas, and up to 95% with the most prominent 524 lemmas.

adjectives, verbs, etc. (see Figure 3d). Our data contains mainly nouns and adjectives, as expected in texts of a descriptive nature.

4 UNDERSTANDING THE VISUAL LANGUAGE OF FABRICS

We conduct a comprehensive analysis of our dataset to understand how people describe fabrics, and explore relevant questions about its characteristics. From these questions, around which this section is structured, we gather insights which help the design of tools to describe, retrieve, classify, label or edit fabrics, among others.

4.1 Is there a common lexicon when describing fabrics?

The existence of a common vocabulary when describing fabrics is a necessary condition for any text-to-fabric application to be practical and successful. Ideally, we would like to identify a reduced set of lemmas or root words (see Section 3.2.3) which would be sufficient for the majority of fabric descriptions.

We begin by computing the *absolute frequency per lemma* $f(w)$ in the full corpus of descriptions, where, for each lemma w, the count includes occurrences of all the single words or types belonging to it. Prominence of lemmas, however, is not only determined by their absolute frequency, but also by their distribution; for instance, if a describer uses words belonging to a lemma often, but other describers do not, the lemma may not be prominent. We therefore complement absolute frequency with a measure of dispersion, indicating how evenly the occurrences of the lemma are distributed within the corpus. We measure this with the *average reduced frequency* ($arf(w)$) [Brezina 2018; Savický and Hlavácová 2002], which modulates $f(w)$ with the dispersion of w (details on the computation can be found in the supplemental material). The histogram of $arf(w)$ (Figure 4 (left)) shows how over one third of the lemmas have an arf value below 2, meaning that they are seldom used, or used by a single describer. This confirms the intuition that a reduced subset of lemmas should suffice for fabric description.

We next examine how small this reduced lexicon can be. We first use $arf(w)$ as ranking criterion to find the subset $\mathcal{W}_k$ of the most prominent k lemmas, for increasing values of $k \in [1..N_w]$, with $N_w = 2,762$ the number of lemmas in our corpus. For each $\mathcal{W}_k$, we then compute the *coverage* of a description d as

$$cov_k(d) = \frac{n_k(d)}{n_{tot}(d)},$$

where n_k is the number of lemmas from subset $\mathcal{W}_k$ present in description d, and $n_{tot}(d)$ is the total number of lemmas of such description. As the plot in Figure 4 (right) shows, a common lexicon of 84 lemmas is enough to cover 75% of the fabric descriptions, while to cover 95% we only need 524 lemmas, which we define as our fabric-specific lexicon (examples of these lemmas can be found in the supplemental material).

4.2 Are there key attributes in the descriptions?

Finding common attributes is useful to understand how we internally represent and think about fabrics. From our reduced 524-lemma lexicon, we seek now to identify common attributes that these lemmas may relate to. We approach this as a clustering problem, and develop a methodology based on affinity propagation and similarity between lemmas. In particular, we leverage embeddings of the lemmas provided by *ConceptNet Numberbatch* [Speer et al. 2017], which combines both distributional semantics and relational

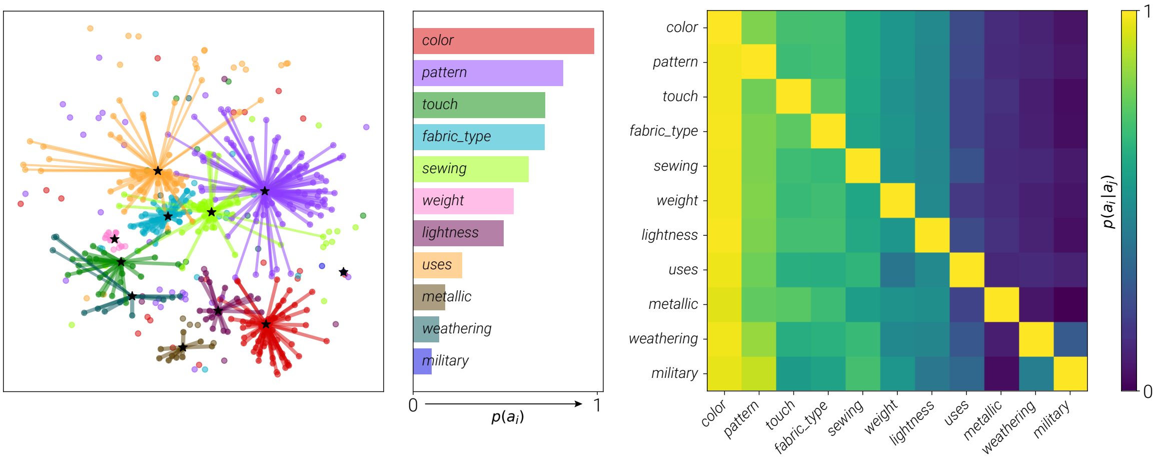


Fig. 5. Attributes present in fabric descriptions. *Left:* Visualization of the lemma embeddings space and its clustering into attributes. We show all lemmas from our lexicon (524) as points in 2D space using t-SNE dimensionality reduction (300D to 2D); the color of every point indicates its associated attribute; we show a line from every non-outlier point to its attribute centroid (marked with a black star). *Center:* Probability of occurrence of each attribute $p(a_i)$, i.e., probability of having at least one occurrence of a lemma belonging to attribute a_i in a description. *Right:* Matrix displaying $p(a_i|a_j)$, with a_i in the columns and a_j in the rows. Note that it is not symmetric because $p(a_i|a_j) \neq p(a_j|a_i)$. We observe how the probability of an attribute appearing is largely independent of the occurrence of other attributes.

knowledge[4]. This leads to the identification of the main attributes that people focus on when describing fabrics, as well as a distribution of our lexicon into those attributes (we provide a description of this process in the supplemental material). This results in eleven key attributes describing fabrics: *color*, *lightness*, *metallic*, *pattern*, *fabric_type*, *sewing*, *touch*, *weight*, *use*, *weathering*, and *military*[5], and are shown in Figure 5 (left) using t-SNE dimensionality reduction [Van der Maaten and Hinton 2008].

In Figure 5 (center), we show the probability of occurrence $p(a_i)$, $i \in [1..N_a]$ of each attribute a_i, where $N_a = 11$ is the number of attributes. It expresses the probability that there is at least one occurrence of a lemma belonging to the attribute in any given description. This illustrates the relative importance of each attribute: for instance, it reveals that *color*, *pattern*, *touch* and *fabric_type* are present in more than 70% of the descriptions.

Moreover, we look into whether certain attributes tend to appear together in the descriptions; to that end we compute $p(a_i|a_j)$, $i, j \in [1..N_a]$, $i \neq j$, i.e., the probability of attribute a_i being present in a description that contains attribute a_j. Figure 5 (right) plots these probabilities for all attributes (note that the resulting matrix is non-symmetric, since $p(a_i|a_j) \neq p(a_j|a_i)$). We observe that, in general, the presence of a given attribute in a description is not heavily dependent on the presence of any other attribute. This is indicated by the relatively uniform values along each column, and is a result of the large variety of appearances present in our dataset, exhibiting many different combinations of attributes.

4.3 Do descriptions follow a common structure?

We next look at the structure of descriptions by analyzing the order of appearance of the different attributes. Specifically, we compute a *rank product* for each attribute as

$$\Psi(a) = \left(\prod_{i=1}^{D} r_{a,i}\right)^{1/D},$$

where $r_{a,i}$ is the *rank* of attribute a in description d_i, $i \in [1..D]$ [Rubinstein et al. 2010]. The rank is given by the first appearance of a lemma belonging to an attribute in a description; thus, lower rank products indicate that the attribute tends to appear earlier in the descriptions.

Table 1 shows the resulting ranking of attributes. To evaluate whether the differences in ordering are significant, we perform a Kruskal-Wallis test (a non-parametric extension of ANOVA, since rankings are an ordinal value and typically cannot be assumed to follow a normal distribution), which shows that there is a significant difference between attributes ($H(10) = 8235.53$, $p < .0001$). A subsequent pairwise comparisons test allows us to identify groups of attributes where there is no significant difference between their mean ranks (also shown in Table 1). The rank histograms per attribute can be found in the supplemental material.

4.4 Does the same fabric elicit similar descriptions?

We measure similarity between descriptions using two state-of-the-art NLP models that have been shown to work well on Semantic Textual Similarity (STS): sentence-T5 [Ni et al. 2021], designed to provide sentence embeddings from text-to-text transformers, and MPNet [Song et al. 2020], shown to work well for semantic search

[4] We use the implementation from https://github.com/commonsense/conceptnet-numberbatch. We refer the reader to the original ConceptNet paper [Liu and Singh 2004], as well the ConceptNet Numberbatch extension [Speer et al. 2017] for more details.

[5] This last attribute reflects the significant amount of samples of a military nature in our dataset and may not generalize to others, see also Section 6.

Table 1. Attributes sorted by rank product, indicative of their order of appearance within a description. Lower rank products indicate that the attribute tends to appear earlier in the descriptions. Attributes grouped together in the table yield no significant difference between their mean ranks.

Attribute a	color	lightness	sewing	metallic	pattern	weight	military	fabric_type	weathering	touch	use
Rank product $\Psi(a)$	2.25	2.39	2.75	2.77	2.86	2.95	3.06	3.17	3.46	3.73	4.25

Table 2. Similarity between descriptions of the same image (intra-image) and descriptions of different images (inter-image). We report, for two state-of-the-art sentence embeddings (sentence-T5 and MPNet): average intra-image and inter-image similarities (and associated standard deviations), test statistics for ANOSIM, and associated p-value (we use a p-value of 0.05 to indicate significance). The descriptions in our dataset exhibit high intra-image similarity, and the statistical test shows that intra-image similarities are significantly larger than the inter-image ones.

	Similarity			
	Intra-image	Inter-image	Test Statistic	p-value
sentence-T5	0.874 (0.037)	0.822 (0.037)	0.694	0.001
MPNet	0.704 (0.087)	0.627 (0.083)	0.497	0.001

using sentence embeddings [Reimers and Gurevych 2019]. Specifically, we compute cosine similarity between the embeddings of our full descriptions, as in the original publications.

To compute the *intra-image* description similarity (similarity between descriptions of the same fabric), we average over all pairwise comparisons in our whole corpus, provided that the two members of a pair belong to the same image. To compute the *inter-image* description similarity (similarity between descriptions of different fabrics), we average over all pairwise comparisons in our whole corpus, provided that the two members of a pair belong to different images.

The results, shown in Table 2, yield a high intra-image similarity (cosine similarities are bounded between -1 and 1), suggesting that the same fabric does indeed elicit similar descriptions by different people. Compared to the inter-image similarity (which one may treat as a baseline), the average intra-image similarity is larger for both models. To test whether these differences are statistically significant, we resort to ANOSIM (analysis of similarities) [Clarke 1993; Warton et al. 2012]. ANOSIM works on all pairwise similarities (or distances) between points (descriptions), and is designed to test the null hypothesis that the similarity between groups (inter-image) is greater than or equal to the similarity within groups (intra-image). We use a p-value of 0.05 to indicate significance and the test statistic as the measure of effect size [Somerfield et al. 2021]. This value is constrained to $[-1, 1]$, with 1 indicating very high intra-image similarity with respect to inter-image similarity, and negative values indicating higher inter-image similarity. Results of this analysis show reasonably high intra-image similarity with respect to inter-image similarity, confirming that the difference is statistically significant (see Table 2).

5 LARGE VISION-LANGUAGE MODEL COUPLING

Our dataset links the appearance of fabrics with natural language, helping to better understand how people describe such materials despite their semantic proximity. Besides, it provides high quality image and associated text data, in large albeit lower amounts than those present in very large-scale datasets used to train recent, very successful vision-language models (see Section 5.1). In this section, we explore applications of our dataset with such models, and to what extent a relatively low amount of high-quality, specialized data improves over their native versions for specific areas such as material appearance. Specifically, we demonstrate text-based fine-grained retrieval, image-based search, and description generation, as well as an improvement of invariance of the image latent representations to light and geometry changes, contributing to, e.g., a more robust notion of appearance similarity. While we show here varied results and evaluations, please also refer to the supplemental material for additional examples.

5.1 Large Vision-Language Models

Recent progress in joint text and image encoding has been enabled by large vision-language models. In this section in particular, we fine-tune and compare to two of the most widely-used models: CLIP [Radford et al. 2021] and BLIP [Li et al. 2022].

CLIP is a neural model composed of two encoders, one for each modality (text and image), which are trained using pairs of text and images. The method relies on contrastive learning [Chen et al. 2020] to encourage encodings of texts and images to lie close to one another in latent space. This has been shown to draw very interesting connections based on the data it is trained on [Goh et al. 2021]. CLIP is particularly powerful thanks to the vast LAION dataset on which it is trained, containing 400 million image-text pairs gathered from the internet. Different encoder architectures have been published, but in this paper we use the ViT-B/16 version, which relies on a visual transformer [Dosovitskiy et al. 2021] with a patch size of 16 × 16.

BLIP is a combination of networks trained jointly, including a pair of encoders, similar to CLIP. Moreover, BLIP also contains a generative head, trained jointly with the rest of the network, enabling it to generate captions corresponding to an image. It is also trained on hundreds of millions of images, including a self-supervised augmentation mechanism called “CapFilt”. Similarly to CLIP, we use the ViT-B version of the network.

While both CLIP and BLIP are trained on very large-scale datasets, the text data to which they are exposed is limited to low quality online captions of images. We will show, in the remainder of this section, that a small amount of high quality data is sufficient to significantly improve the networks’ sensitivities to specialised concepts. In the following experiments, we use the models published by SalesForce and OpenAI.

5.2 Text-Based Fine-Grained Retrieval

Given a query in the form of a text description, the goal of this first application is to retrieve fabric samples that match the query. Improving search performance in large datasets is increasingly important as the number and size of libraries and datasets increase [Quixel Megascans 2023; Substance 3D Assets 2023]. Different from generic text-based retrieval, which aims at finding images of objects in different classes such as “chairs”, “cars”, or “people”, we target the more difficult case of *fine-grained* retrieval [Qi et al. 2021], i.e., finding the right instance despite significant semantic similarities within the dataset.

5.2.1 Implementation. We fine-tune CLIP [Radford et al. 2021], starting from the VIT-B/16 pre-trained model (we term this pre-trained model *native* CLIP). Our dataset is split in 12,334 training and 3,129 test descriptions, ensuring that no procedural variation of a given material in the training data is used for testing. During training we split the descriptions by sentence, resulting in 45,871 (36,565 train/9,306 test) individual sentence descriptions for 3,000 (2,393/607) materials. We train the network for 12 epochs (at which point the performance levels off, with small gains until epoch 19), using renderings of the training materials on four geometries (*baseline*, *sphere*, *sphere_draped*, *plane*) and a single illumination (*baseline*). We use a learning rate of $1e^{-6}$ with a linear schedule with 200 warm-up steps for the Adam optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and batch size 128. This takes five hours to train on a single Nvidia RTX3090 GPU. For inference, execution time is 0.46 seconds for a batch of 64 images.

5.2.2 Results. Since we have ground-truth data (image-description pairs in our test dataset), we can evaluate retrieval of the *correct* material given an input description. Additionally, for any given generic query, the retrieval application should return relevant results. The search operation presented in this section is made over the entire 3,000 materials on our *baseline* geometry and illumination unless specified otherwise; in all cases, neither the descriptions used as queries nor the correct images or materials have been seen during training.

Quantitative analysis. Given our test set descriptions, we evaluate retrieval in the complete material database and report the top-K recall results of our fine-tuned CLIP, with $K \in \{1, 5, 10, 20, 100\}$, in Table 3. We also include results for *native* CLIP, *native* BLIP[6], and BLIP trained on our data only (BLIP *no pretrain*). Compared to native CLIP/BLIP, we achieve 4.8/4 times better top-1 retrieval rate and maintain at least 2.12 times better results for all top-K results, showing that our dataset makes CLIP more sensitive to fabric-specific concepts. This also shows that our fine-tuned model is capable of retrieving a fabric sample from its description alone, which requires strong feature discrimination in a semantically similar dataset. The comparison to BLIP trained on our data only (no pretrain) shows that our model significantly benefits from the original model training, leveraging the priors provided by its large corpus of text.

To evaluate the ability of our model to generalize to other geometries, Table 4 reports retrieval recall results on a geometry unseen during training (*plane_draped*). We see that, despite the unseen geometry being challenging (all methods have lower retrieval results than with the *baseline* geometry), our fine-tuned model still significantly benefits from our dataset. Furthermore, we also include a comparison to our model fine-tuned on images from just one geometry (*baseline*), showing that the training on different geometries improves the generalization of the method.

Table 3. Top-K retrieval results on the *baseline* geometry for native CLIP, native BLIP, BLIP trained on our data only (BLIP no pretrain) and our fine-tuned model.

	Native CLIP	Native BLIP	BLIP no pretrain	Ours
Top-1	2.94%	3.42%	1.60%	**13.81%**
Top-5	8.31%	9.94%	5.98%	**33.91%**
Top-10	12.59%	14.60%	10.64%	**46.76%**
Top-20	18.37%	20.17%	17.00%	**59.76%**
Top-100	41.29%	34.26%	34.36%	**87.63%**

Table 4. Top-K retrieval results on the *plane_draped* geometry, unseen during training, for native CLIP, native BLIP, our model fine-tuned on only one geometry (*baseline*), and our model (which is fine-tuned on four geometries, not including *plane_draped*).

	Native CLIP	Native BLIP	Ours (1 gm.)	Ours (4 gm.)
Top-1	1.34%	2.27%	5.98%	**7.38%**
Top-5	5.02%	7.03%	16.55%	**22.95%**
Top-10	7.93%	10.9%	23.94%	**31.77%**
Top-20	12.66%	15.05%	34.16%	**43.72%**
Top-100	32.18%	29.05%	64.59%	**75.93%**

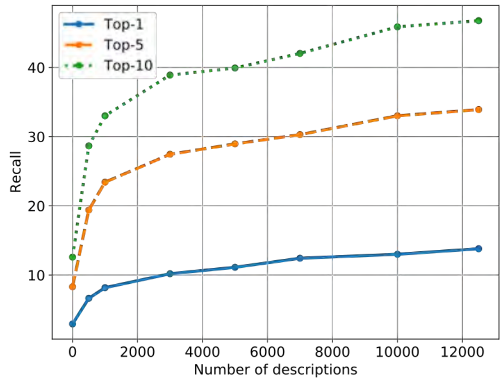


Fig. 6. Evolution of text-based retrieval results (top-1, top-5 and top-10 recall performance) with the number of descriptions available for fine-tuning. Native CLIP performance corresponds to the case of zero descriptions available for fine-tuning.

We also assess the required size of a specialized dataset such as ours for fine-tuning general purpose models. To do so, we plot the top-K retrieval results as we vary the number of descriptions used in the training in Figure 6. We see how the model significantly benefits from the first 2,000 descriptions, and how the marginal improvement rate then starts diminishing. At constant number of descriptions, we also evaluate whether more images with fewer descriptions is preferable to fewer images with more descriptions. We find that using 1,500 images with 5 descriptions per image is equivalent in retrieval recall to using 2,500 images with 3 descriptions per image: Results in both cases are close, indicating a similar impact between image and description diversity. More precisely, 1,500 images with 5 descriptions each yield top-1/5/10 retrieval results of 12.56/31.38/44.1%; in comparison, 2,500 images with 3 descriptions each yield top-1/5/10 retrieval results of 12.66/31.16/43.69%. In addition, we also include

[6] For details on the implementation of native BLIP please refer to Section 5.4.1.

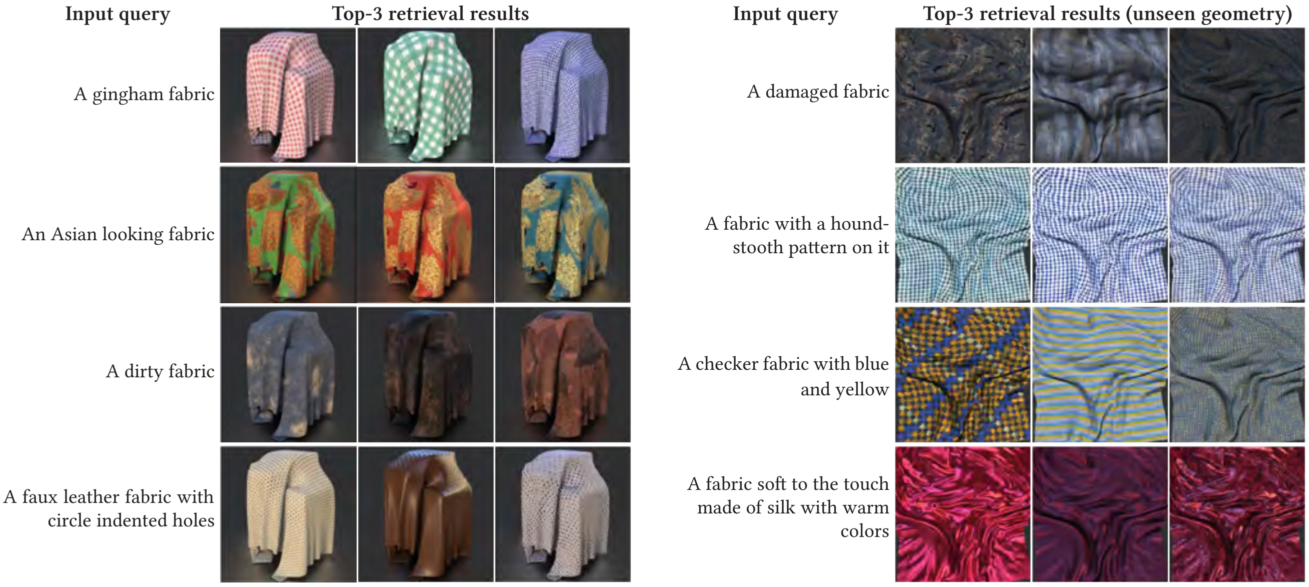

Input query
Top-3 retrieval results
A gingham fabric
An Asian looking fabric
A dirty fabric
A faux leather fabric with circle indented holes
Input query
Top-3 retrieval results (unseen geometry)
A damaged fabric
A fabric with a hound-stooth pattern on it
A checker fabric with blue and yellow
A fabric soft to the touch made of silk with warm colors

Fig. 7. Text-based fine-grained retrieval, evaluating the sensitivity of our fine-tuned representation to varied domain-specific concepts on two different geometries. We show input text queries, and the top-3 retrieval results using our fine-tuned model. *Left:* Retrieval results on the *baseline* geometry, seen during fine-tuning. *Right:* Retrieval results on the *plane_draped* geometry, unseen during fine-tuning. Our model retrieves relevant results for aspects related to different attributes, and for both high-level and more specific queries.

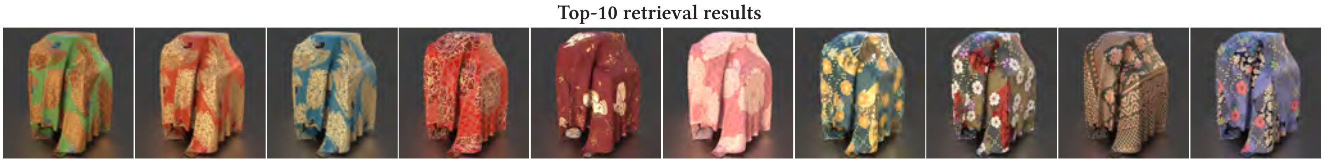


Fig. 8. Top-10 results of text-based fine-grained retrieval with our fine-tuned model for the query "An Asian looking fabric". Although the closest samples (top-3, also shown in Figure 7) have similar appearance, we observe more diverse results (while still relevant) when increasing the number of images returned.

a quantitative evaluation on negative queries in the supplemental material.

Qualitative analysis. To qualitatively evaluate the performance of our fine-tuned model, we provide results retrieved from natural language queries in Figure 7, showing that the retrieved materials exhibit the desired properties, not only in a geometry seen during training[7] (*baseline*), but also in unseen geometry (*plane_draped*). As expected, diversity increases as we look at more returned samples. This can be seen in our "Asian looking" prompt; while the top-3 results in Figure 7 contain similar results due to our space being partly organized with respect to visual features, more diverse results appear when visualizing the top-10 results, as shown in Figure 8.

In Figure 9, we show a more systematic evaluation with positive and negative queries, corresponding to prominent fabrics concepts extracted from the dataset (see Section 4), and include a comparison to native CLIP. Results confirm that our fine-tuned model is effective in the retrieval, and more sensitive to fine-grained descriptions, while native CLIP struggles with specialized concepts (e.g., stitching) and negative wording. Interestingly, despite the relatively small amount of data used in our fine-tuning (compared to the hundreds of millions of image-text pairs required to train CLIP and BLIP), we observe significant improvement in material retrieval for the class of interest (fabrics). Furthermore, these experiments highlight the limitations existing in the representations of Large Vision-Language Models for fine-grained appearance concepts.

Finally, we evaluate the limits of modeling out-of-distribution queries, containing concepts that do not appear in our dataset descriptions. As shown in Figure 10, our model finds reasonable results for these queries (e.g., for the case of "Thanksgiving-themed" we obtain a variety of autumnal brown and orange fabrics), suggesting that our model preserves its broader priors without overfitting to our dataset.

5.3 Image-Based Search

We continue our evaluation by studying image-based search using real images as input. We do it by leveraging our fine-tuned CLIP model (see Section 5.2), as well as native CLIP for comparison. Specifically, we compute the normalized embedding—using either native

[7] Note that the geometry has been seen during training, but the materials and queries have not.

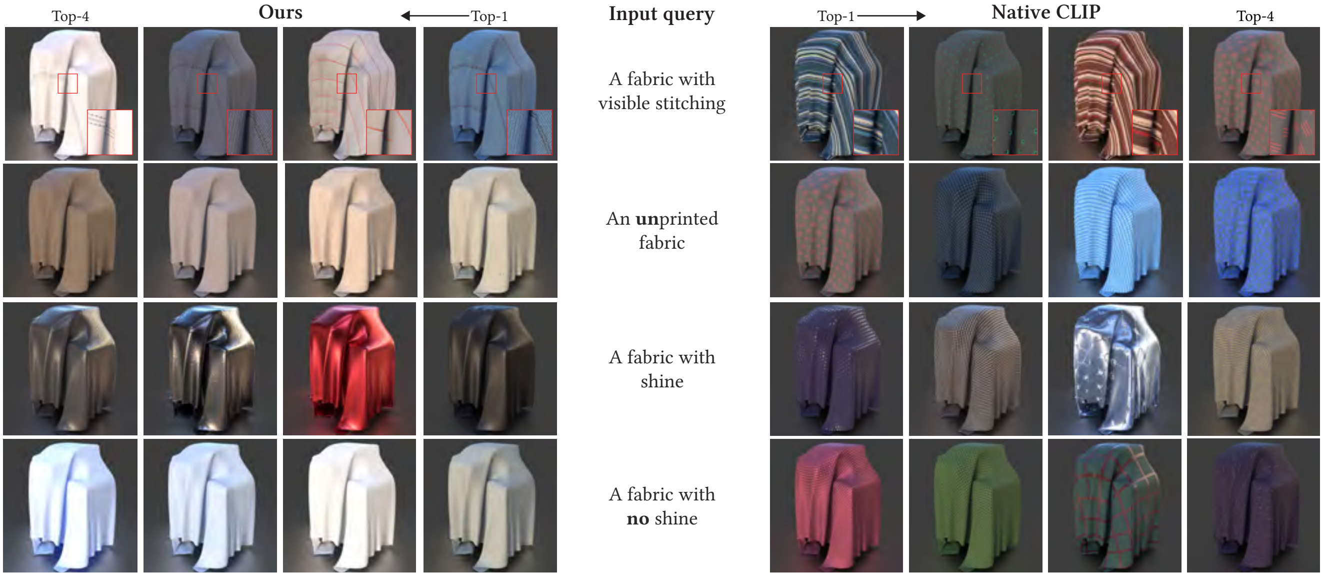


Fig. 9. Text-based fine-grained retrieval. We evaluate the sensitivity of our fine-tuned model to different text queries, including negative ones, and compare it with native CLIP. Our model is more sensitive to domain-specific concepts and can better handle negative queries.

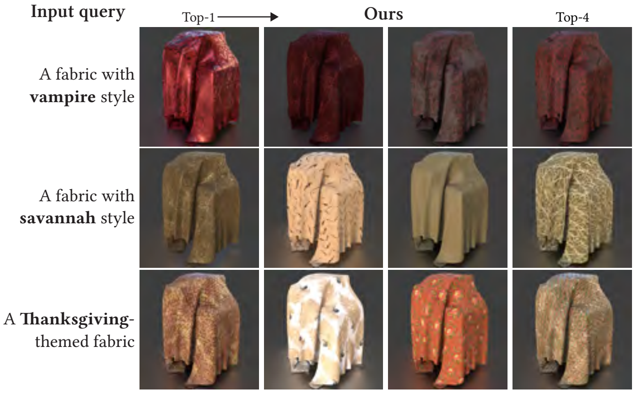

Input query
Top-1
Ours
Top-4
A fabric with **vampire** style
A fabric with **savannah** style
A **Thanksgiving**-themed fabric

Fig. 10. Text-based fine-grained retrieval with out-of-distribution queries that contain concepts not included in our dataset descriptions (marked in bold). We show top-4 retrieval results using our fine-tuned model.

CLIP or our fine-tuned model—of the input image, and compute its cosine distance to the normalized embeddings of the candidates from our dataset. These candidates are the 3,000 materials in our dataset, rendered on a certain geometry (or set of geometries in Section 5.5). Figure 11 shows results on the *plane_draped* geometry (unseen during training) for both our fine-tuned model and native CLIP. We can observe that native CLIP is strongly influenced by the geometrical macrostructure present in the input image, and fails at guiding the retrieval process by the material mesostructure, patterns and reflectivity properties expressed in the input. On the contrary, our fine-tuned model succeeds at fetching results with similarities existing at material scale, bypassing the strong features stemming from the supporting 3D shape.

5.4 Caption Generation

Caption generation aims at creating accurate descriptions of a fabric material given an image of it. Similarly to the retrieval application, we target fine-grained description, with precise properties described, which are not limited to high-level semantics. This further allows us to explicitly observe the ingestion of the concepts stemming from our dataset by Large Language Models.

5.4.1 Implementation. We leverage and fine-tune BLIP [Li et al. 2022] for caption generation, and process our data as described in Section 5.2.1; however, in this case we do not split the sentences, ensuring full descriptions are seen by the model. We fine-tune the generative head of BLIP, starting from what we term *native* BLIP: the VIT-B/16 model pre-trained on 129M images from LAION + CapFilt-L (model_base_capfilt_large). We train the network for 12 epochs using the Adam optimizer with weight decay regularization using a decay parameter of 0.05, an initial learning rate of $1e^{-5}$ and a batch size of 24. The minimum number of generated tokens is set to 5, and the maximum to 80. This takes approximately 9 hours to train on a single RTX3090. Once fine-tuned, we use nucleus sampling for tokens [Holtzman et al. 2020], letting us generate varied descriptions for each image.

5.4.2 Keyword Extraction. While we mainly focus on generating natural language descriptions (using the model described in Section 5.4.1), simple keywords can also be convenient in several search or classification scenarios. Therefore, using our understanding of fabric descriptions in terms of common lexicon, main attributes and structure (Section 4), we automatically extract keywords for an image based on the generated descriptions. For that, we use the first five descriptions of an image, post-process the text as explained in Section 3.2.3, extract a set of keywords per attribute from our lexicon, and order them by importance (number of descriptions in

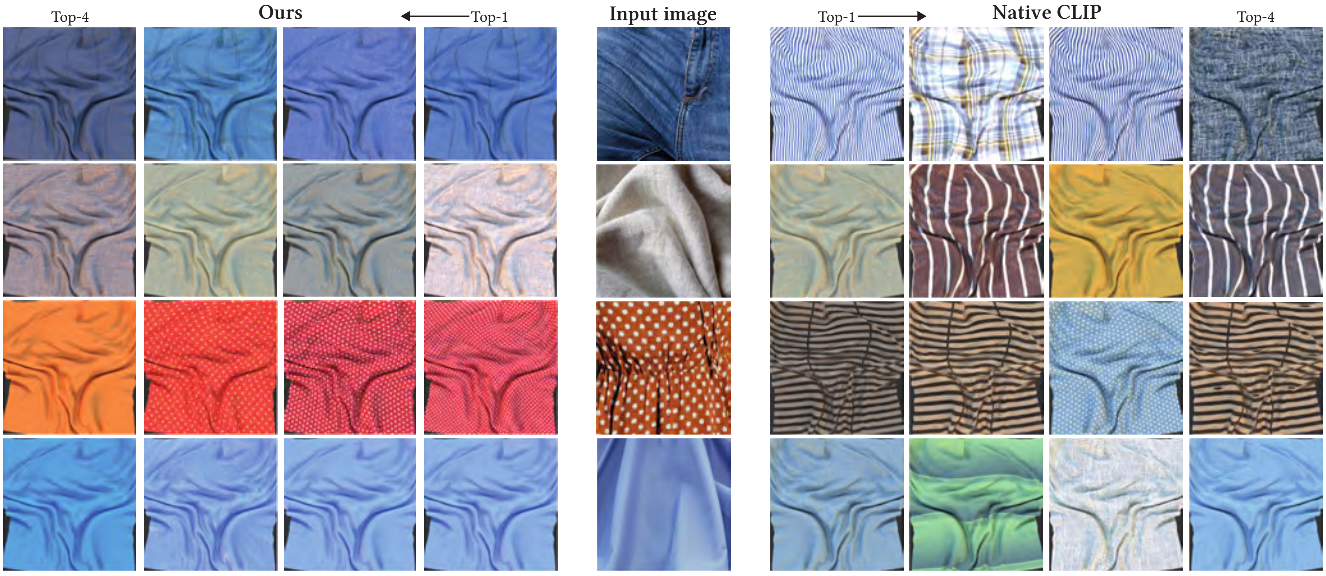


Fig. 11. Image-based search with real photographs as input (*middle column*), using our fine-tuned model (*left*) and using native CLIP (*right*). The search is performed on image data from a single geometry (*plane_draped*), unseen during the fine-tuning. Our model, unlike native CLIP, is capable of retrieving results that are in close correspondence with the input, while circumventing the prominent characteristics arising from the macroscopic geometric structure.

which they appear) and the rank product of their attribute. Resulting keywords for real images are shown in Figure 13, and we include automatically extracted keywords for all material samples as part of our text2fabric dataset.

5.4.3 Results. We show here description results from our fine-tuned BLIP model, together with comparisons to native BLIP.

Figure 12 shows captioning results on synthetic images, with two different geometries, and materials from the test set (unseen during training). For each fabric sample, we include: descriptions from our dataset, provided by humans; descriptions generated by our fine-tuned model; and descriptions generated by native BLIP. We observe how our fine-tuned model generates descriptions that better convey fine-grained material appearance, are more accurate and with more attention to detail, and match more closely the style of human descriptions.

We further demonstrate our results on real images containing fabrics in Figure 13. We crop the fabric area of interest (marked by a red square), and generate descriptions for it. The descriptions generated using our fine-tuned model contain significantly richer information than the native BLIP results, trained only on general internet images and high-level descriptions. Additionally, our keyword extraction method is capable of automatically extracting relevant keywords from our generated sentences, which can be useful for, e.g., automatic tagging. These results further show that the fine-tuning on our high quality renderings generalizes well to real photographs.

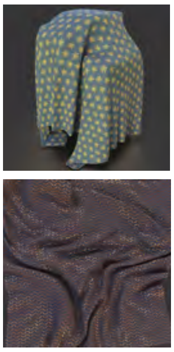
5.5 Invariance of the Latent Space to Geometry and Illumination

Our dataset significantly helps to improve the invariance to lighting and geometry of large vision-language models representations. We illustrate this by fine-tuning CLIP (as described in Section 5.2.1) using all the renderings of our materials, with five different geometries and three environment illuminations (*baseline*, *outdoor*, *studio*) associated to our 15,461 descriptions. While we could use contrastive learning to try to learn an invariant representation instead, the only available supervision would be whether or not two images show the same material, making the creation of a perceptually smooth representation challenging. Using our descriptions as anchor contributes to a smooth representation space, allowing the model to learn a more robust notion of material appearance than that of native CLIP, as shown by our evaluation, described next.

Table 5. Average cosine similarity (and associated standard deviation) between pairs of images exhibiting: the same material (and lighting conditions) and different geometries (*varying geometry*); and the same material (and geometry) and different lighting conditions (*varying lighting*). Our fine-tuned representation (*first row*) finds images of the same material to be more similar, despite geometry or lighting variation, than native CLIP space (*second row*). We include results of the Wilcoxon signed-rank test showing effect size | p-value (*bottom row*). Differences are statistically significant and with large effect sizes for both experiments.

	Varying Geometry	Varying Lighting
Ours	0.951 ± 0.012	0.973 ± 0.007
Native CLIP	0.835 ± 0.042	0.945 ± 0.011
Wilcoxon signed-rank	0.866 | <0.0001	0.846 | <0.0001

In Table 5 (first column, *varying geometry*) we evaluate the average cosine similarity between pairs of images rendered with the same material and lighting but different geometries, computed both in native CLIP space and in the latent space of our fine-tuned model. Our representation is, on average, more invariant to geometry than the original CLIP features. The same evaluation for



Input	Human description	Ours	Native BLIP
	This fabric resembles canvas and is woven with black and blue yarn giving it a dull gray appearance from afar. It has a slightly rough texture and no sheen. There are golden sunbursts that appear to made with fabric paint.	This fabric has a blue background with small yellow suns all over it the sun is designed in repeating lines of five - point stars the linen thread could be used to make curtains or upholster	This is a print of gold flowers in an ocean blue background, and placed over the black
	This is blue cotton fabric. It is made up of yellow colored sunlight patterns. it appears to be soft and is mostly used for cushions and table linen.	This fabric is a muted gold / pale yellow on a slate blue it has an appearance of being woven, like a sun with five - point star patterns similar to the stars themselves	The cover is made out of blue and yellow flowers
	This is a medium weight fabric. The base color is a coffee brown. The overall pattern is continual broken zigzag lines of copper orange throughout.	This fabric is black in color, with a pattern of metallic gold zigzags these are arranged diagonally throughout the material it looks tightly knit and soft to the touch	The navy blue chey print on this tie is made of silk and has a thin pattern
	This fabric is dark gray with zigzag designs throughout the fabric in orange. Its texture is soft and smooth. The material seems most likely to be polyester. The orange zigzag features a little sparkle.	This fabric is dark gray with a pattern consisting of many jagged triangles throughout these triangular shapes are gold the material appears shiny and smooth, even satin it would make an elegant dress	A blue scarf with orange pattern on it

Fig. 12. Description generation results for synthetic materials from the test set. We show results on two different geometries: *baseline* (*top*), seen during fine-tuning, and *plane_draped* (*bottom*), unseen during fine-tuning. The descriptions included are, *from left to right*: from our gathered dataset, provided by humans; generated by our fine-tuned model; and generated by native BLIP. Our descriptions not only are closer in style to human descriptions, but are also better at conveying fine-grained appearance and details.

Input	Ours (keywords)	Ours	Native BLIP
	Yellow, Bold, Stitch, Shiny, Gold, Horizontal, Lightweight, Silk, Smooth	This fabric has a lot of shine to it and could be silk with a slick looking feel to it the color is all gold shiny and looks lightweight	The front of a gold and black gown, which is made from silk
		The fabric is a shiny solid yellow that has no obvious shading or patterns in it the material is probably made out of a polyester or similar and would be lightweight	Mustard fabric, satin satine poly span nylon
	Brown, White, Stitch, Weave, Lightweight, Thick, Wool, Polyester, Rough, Texture	This fabric looks like a cotton or linen material with stitching on the surface and colors of grey and white all mixed together this looks smooth and lightweight	Linen fabric closeup texture white
		The fabric is a twill weave with brown and cream yarns it is fairly thick and appears rough to the touch the fibre content could be wool, cotton or polyester	Natural linen fabric - cotton & cashmere

Fig. 13. Description generation results on real images (the input is marked by a red square) for both our fine-tuned model and native BLIP. We also include results of our automatic keyword extraction. We can see that our fine-tuned model for caption generation generalizes well to real photographs.

pairs of images rendered with the same material and geometry but different lighting conditions (Table 5, *varying lighting*), shows a similar trend, although less pronounced. In both cases, the lower standard deviation between pairwise similarities using our representation suggests a greater stability across variations. A Wilcoxon signed-rank test shows that these differences are statistically significant (p-value<0.0001), with effect sizes considered large for both geometry and lighting variations [Rosenthal et al. 1994].

We qualitatively evaluate this property in Figure 14: we assess whether, given a real photograph as input to an image-based search (see Section 5.3), the results change depending on the geometry present in the database we search in (for this test, each database we search in has all materials rendered with a single geometry). The figure shows results for the search in the *sphere_draped* and *plane* geometries databases, and we display all results rendered on *sphere_draped* for easier comparison. We can see that our representation is significantly more consistent than native CLIP on varying geometries, and better at learning features at material scale.

We further pursue this evaluation in Figure 15. Here, we seek at retrieving a given *test* (i.e., unseen during training) material rendered on a given geometry, performing image-based search in a database containing renderings of all 3,000 materials and five geometries. As expected, with both native CLIP and our fine-tuned model, the first result is the same material and geometry. However, it is clearly apparent that the native CLIP representation is heavily biased by the geometry in the input image, while our representation better identifies the same material across geometries.

6 DISCUSSION AND FUTURE WORK

We have presented text2fabric, a comprehensive, large-scale public dataset relating the visual appearance of fabrics to natural language. We have analyzed and curated a rich lexicon, classifying it into eleven attributes and highlighting key concepts used by humans when describing fabrics. We have further proposed several applications including fine-grained retrieval, image-based search, and caption generation, and shown how foundational, state-of-the-art vision-language models such as CLIP [Radford et al. 2021] or BLIP [Li et al. 2022] struggle to represent fine-grained concepts of appearance, unless fine-tuned on our dataset.

Our work is not free of limitations. First, as in all studies, our results are only strictly valid for our particular set of stimuli; for example, in the fabric samples used to generate our data, military characteristics and some weathering features are highly correlated,

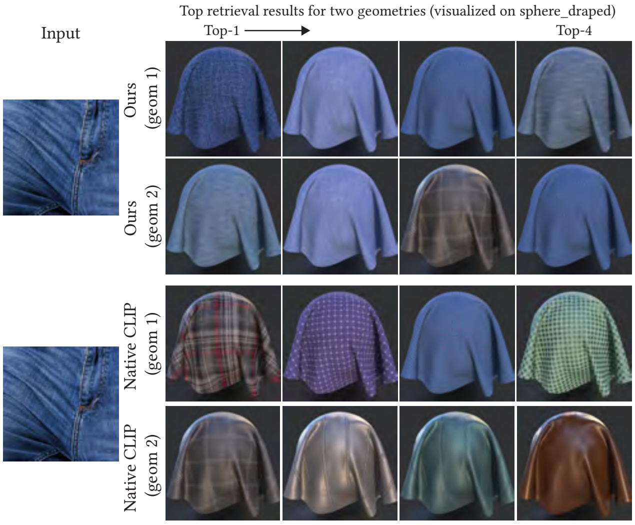


Fig. 14. Latent space invariance to geometry. Top-4 results of image-based search in databases rendered on different geometries (geom 1: *sphere_draped*; geom 2: *plane*; see text for details). We display all results rendered on *sphere_draped* for easier comparison. Our representation is significantly less affected by the geometry than the latent space of native CLIP, learning a more precise notion of material appearance.

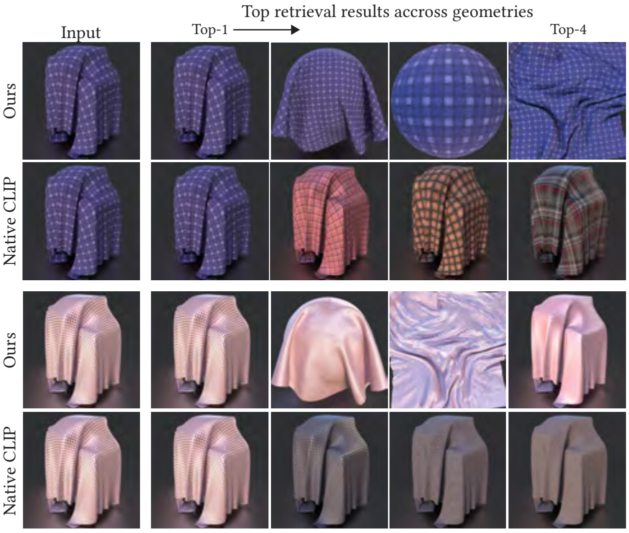
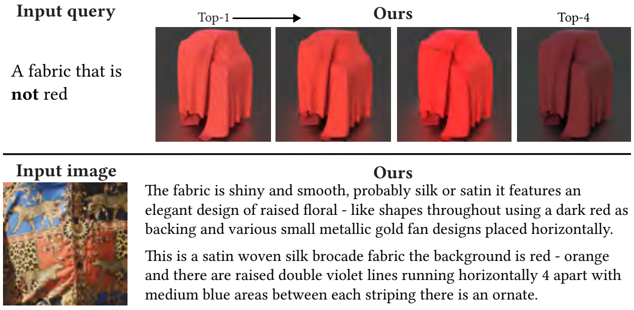


Fig. 15. Top-4 results of image-based search in a database containing all our geometries: We see that native CLIP is heavily biased by the geometry of the input image, while our fine-tuned model focuses on the material appearance and is capable of recovering the same material across geometries.

such as camouflage and dirt, as shown in Figure 5 (right). This leaves the door open for future extensions of our dataset to explore these correlations further. Second, we decided to choose non-expert describers (albeit familiar with fashion or design), to target a wider audience for our applications, given that experts usually rely on highly specialized concepts, difficult to understand by the general public. This might lead to the descriptions including some inaccuracies



The fabric is shiny and smooth, probably silk or satin it features an elegant design of raised floral - like shapes throughout using a dark red as backing and various small metallic gold fan designs placed horizontally.

This is a satin woven silk brocade fabric the background is red - orange and there are raised double violet lines running horizontally 4 apart with medium blue areas between each striping there is an ornate.

Fig. 16. Limitations. *Top row:* The text-based fine-grained retrieval does not always work well for negative queries that people do not use when describing (e.g., one would not say "this is a non-red fabric" or "this fabric is not red"). *Bottom row:* While our generated descriptions capture many relevant details, the intricacy of the pattern image results in the model missing some features that are salient to humans, such as the leopards.

(e.g., due to uncertainty in the fabric type), or common misunderstandings about cloth (e.g., confusing "stitching" with "weaving"). While these are a reflection of assumptions and biases from common users, it could be a limitation in certain scenarios, e.g., involving experts. Additionally, as expected, some text-based queries are not fully understood by our models, as shown in Figure 16 (top row). Fabric samples are typically not described in terms of *not* having a certain characteristic (people do not say "this is not a red fabric" or "this fabric is not red"); as a result, our fine-tuned models (and their native counterparts) struggle with such queries. Another limitation regarding caption generation occurs in the presence of very complex, intricate designs, where the descriptions produced may fail to capture certain aspects that would be prominent for a human. An example can be seen in Figure 16 (bottom row), where despite the richness of the generated caption, it fails to mention the presence of leopards in the fabric.

Our work opens up exciting avenues for future research, which we describe in the following paragraphs.

Generalization. An interesting question from our work is the exploration of how well our methodology generalizes beyond fabrics, maintaining a similar intra-class variation description quality. We argue that our methodology can be readily applied to other material datasets and classes, including both data gathering and analysis, which could in turn enable similar applications to the ones described in Section 5. As a proof-of-concept, and without incurring in the cost of gathering a whole new corpus of descriptions, we resort to Adobe Stock [2023], a popular service where assets can be searched by class, and are tagged with keywords provided by artists. We gather the keywords corresponding to four material classes ("wood", "stone", "brick" and "metal"), quite different in nature from fabrics; while Adobe Stock does not provide free text descriptions, we aim to assess to what extent the keywords are well represented by our attributes found in Section 4.2.

We remove the attributes that are specific to fabrics, namely: *sewing*, *weight* and *military*, and rename *fabric_type* with the corresponding material category (e.g., *wood_type*). We then automatically

Table 6. Precision (ratio between true positives and predicted positives) of the automatic classification of keywords from other material categories into our generic attributes.

	color	lightness	metallic	pattern	touch	use	weath.	mat_type
wood	0.71	1.00	1.00	0.68	0.75	0.84	1.00	0.89
brick	0.72	0.80	0.67	0.59	0.68	0.83	1.00	0.75
stone	0.52	0.95	0.65	0.52	0.75	0.85	0.83	0.95
metal	0.53	0.47	0.70	0.68	0.75	0.88	1.00	0.67
avg	0.62	0.80	0.76	0.62	0.73	0.85	0.96	0.82

classify keywords from all four classes into the attributes (see Section 4.2). Table 6 shows precision values for each attribute and class, i.e., how many keywords assigned to the attribute truly belong to it (we obtain the ground truth by manual classification). We see how precision values are reasonably high, suggesting generality of our attributes. The exceptions are *color*, whose low precision is due to the presence in our lexicon of common objects used as colors (e.g., olive), and *pattern*, probably due to the very general nature of this attribute. While this is a very preliminary analysis, we believe it hints at the generalization capabilities of our methodology and derived attributes, and may inspire future work in this direction.

Another interesting avenue of research is exploring generalization *beyond* material categories, such as video or meshes. Besides, since our methodology lets us relate synthetic graphics primitives to natural language, we are then free to use our primitives under arbitrary conditions, for example adapting them to a specialized context such as garments, or specific environments. An interesting line of future work would be to exponentially augment datasets by combining geometries and materials descriptions into new complete descriptions of the combination, enabling virtually infinite geometry, environment and material combinations for downstream natural language and visual tasks.

Dataset extension. We used rendered images instead of photographs due to the large size of our dataset, since capturing 45,000 samples of different fabrics under controlled, professional conditions would impose non-negligible costs. On the other hand, using existing photographs would introduce uncontrolled variations in geometry and lighting, which may hamper the task of describing material appearance. Nevertheless, carefully augmenting our dataset with real images could enhance the performance of some applications. Additionally, our dataset could be further extended by adding expert terminology to the textual data, and used for instance to investigate social associations typically derived from clothing, such as occupation, personality, or socioeconomic status.

Generative models. While the challenging task of material generation is out of the scope of this study, recent material generation models have used different images as conditions [Guo et al. 2020; Zhou et al. 2022]. As shown, our dataset enables better visual correspondence between appearance and natural language. Combined with the strong prior of a fabric material generation model, our dataset could significantly improve text-conditioned material generation and editing.

Physical properties. Our dataset consists only of static stimuli. Although it has been shown that visual appearance dominates over dynamics when describing most fabrics, certain characteristics may be better conveyed by simulating the physics of such fabrics in motion [Aliaga et al. 2015]. Exploring the relative weights of appearance and dynamics on the perception of fabrics is an interesting research topic, although requiring a significant amount of work to model and simulate the physics of the fabrics.

We hope that text2fabric helps enable these and other studies, which in turn may lead to the creation of novel applications.

ACKNOWLEDGMENTS

We thank Kushal Kafle for his help in navigating the NLP domain; Yulia Gryaditskaya and Zoya Bylinskii for advice about data collection and user studies; Jorge Gracia del Rio for NLP pointers; and Sergio Izquierdo and the members of the Graphics and Imaging Lab for discussions and help preparing the final figures. We would also like to thank the Substance 3D Assets team for the creation of the materials, all the participants in the user studies, and the anonymous reviewers. This work has received funding from the European Research Council (ERC) within the EU's Horizon 2020 research and innovation programme (project CHAMELEON, Grant No. 682080) and under the Marie Skłodowska-Curie grant agreement No. 956585 (project PRIME), as well as a generous donation from Adobe. Julia Guerrero-Viu was supported by the FPU20/02340 predoctoral grant.

REFERENCES

2001. *Complete Textile Glossary.* Celanese Acetate LLC.

Adobe Stock. 2023. https://stock.adobe.com/.

Carlos Aliaga, Carol O'Sullivan, Diego Gutierrez, and Rasmus Tamstorf. 2015. Sackcloth or silk? The impact of appearance vs dynamics on the perception of animated cloth. *Proc. of the ACM Symposium on Applied Perception* (2015), 41–46.

Dirk Bartz, Douglas W Cunningham, Jan Fischer, and Christian Wallraven. 2008. The Role of Perception for Computer Graphics. *Eurographics (State of the Art Reports)* (2008), 59–80.

Nalini Bhushan, A Ravishankar Rao, and Gerald L Lohse. 1997. The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images. *Cognitive science* 21, 2 (1997), 219–246.

Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. 2011. Describing people: A poselet-based approach to attribute classification. *IEEE International Conference on Computer Vision* (2011), 1543–1550.

Adrien Bousseau, Emmanuelle Chapoulie, Ravi Ramamoorthi, and Maneesh Agrawala. 2011. Optimizing environment maps for material depiction. *Computer Graphics Forum* 30, 4 (2011), 1171–1180.

Ivaylo Boyadzhiev, Kavita Bala, Sylvain Paris, and Edward Adelson. 2015. Band-sifting decomposition for image-based material editing. *ACM Trans. Graph.* 34, 5 (2015), 1–16.

Vaclav Brezina. 2018. *Statistics in corpus linguistics: A practical guide.* Cambridge University Press.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning* (2020), 1597–1607.

Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. 2022. Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *arXiv preprint arXiv:2210.11277* (2022).

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. *IEEE Conference on Computer Vision and Pattern Recognition* (2014), 3606–3613.

K Robert Clarke. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology* 18, 1 (1993), 117–143.

Forrester Cole, Aleksey Golovinskiy, Alex Limpaecher, Heather Stoddart Barros, Adam Finkelstein, Thomas Funkhouser, and Szymon Rusinkiewicz. 2008. Where Do People Draw Lines? *ACM Trans. Graph.* 27, 3 (2008).

Johanna Delanoy, Manuel Lagunas, Jorge Condor, Diego Gutierrez, and Belen Masia. 2022. A Generative Framework for Image-based Editing of Material Appearance using Perceptual Attributes. *Computer Graphics Forum* 41, 1 (2022), 453–464.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations* (2021).

Ron O Dror, Edward H Adelson, and Alan S Willsky. 2001. Estimating surface reflectance properties from images under unknown illumination. *Human Vision and Electronic Imaging VI* 4299 (2001), 231–242.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

Roland W Fleming. 2017. Material Perception. *Annual Review of Vision Science* 3 (2017), 365–388.

Roland W Fleming, Shin'ya Nishida, and Karl R Gegenfurtner. 2015. Perception of material properties. *Vision Research* 115 (2015), 157–162.

Elena Garces, Aseem Agarwala, Diego Gutierrez, and Aaron Hertzmann. 2014. A similarity measure for illustration style. *ACM Trans. Graph.* 33, 4 (2014), 1–9.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. *Distill* (2021).

Yulia Gryaditskaya, Mark Sypesteyn, Jan Willem Hoftijzer, Sylvia Pont, Fredo Durand, and Adrien Bousseau. 2019. OpenSketch: A Richly-Annotated Dataset of Product Design Sketches. *ACM Trans. Graph.* 38, 6 (2019), 232.

Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. *ACM Trans. Graph.* 39, 6 (2020).

Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3, 6 (1973), 610–621.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *International Conference on Learning Representations* (2020).

Adrian Jarabo, Belen Masia, Adrien Bousseau, Fabio Pellacini, and Diego Gutierrez. 2014. How do people edit light fields. *ACM Trans. Graph.* 33, 4 (2014).

Erum Arif Khan, Erik Reinhard, Roland W Fleming, and Heinrich H Bülthoff. 2006. Image-based material editing. *ACM Trans. Graph.* 25, 3 (2006), 654–663.

Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. 2011. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 10 (2011), 1962–1977.

Manuel Lagunas, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. 2019. A Similarity Measure for Material Appearance. *ACM Trans. Graph.* 38, 4 (2019).

Manuel Lagunas, Ana Serrano, Diego Gutierrez, and Belen Masia. 2021. The joint role of geometry and illumination on material recognition. *Journal of Vision* 21, 2 (2021).

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *International Conference on Machine Learning* (2022).

Zhengqi Li and Noah Snavely. 2018. Learning intrinsic image decomposition from watching the world. *IEEE Conference on Computer Vision and Pattern Recognition* (2018), 9039–9048.

Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.

Ann McNamara, Katerina Mania, and Diego Gutierrez. 2011. Perception in graphics, visualization, virtual environments and animation. In *SIGGRAPH Asia 2011 Courses.*

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877* (2021).

Jannik Boll Nielsen, Henrik Wann Jensen, and Ravi Ramamoorthi. 2015. On Optimal, Minimal BRDF Sampling for Reflectance Acquisition. *ACM Trans. Graph.* 34, 6 (2015).

Genevieve Patterson and James Hays. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. *IEEE Conference on Computer Vision and Pattern Recognition* (2012), 2751–2758.

Fabio Pellacini, James A Ferwerda, and Donald P Greenberg. 2000. Toward a psychophysically-based light reflection model for image synthesis. *Proceedings of ACM SIGGRAPH* (2000), 55–64.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Songe. 2021. Towards Fine-Grained Sketch-Based 3D Shape Retrieval. *IEEE Transactions on Image Processing* (2021).

Quixel Megascans. 2023. https://quixel.com/megascans/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning* (2021), 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of Conference on Empirical Methods in Natural Language Processing* (2019).

Robert Rosenthal, Harris Cooper, Larry Hedges, et al. 1994. Parametric measures of effect size. *The handbook of research synthesis* 621, 2 (1994), 231–244.

Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. 2010. A comparative study of image retargeting. *ACM Trans. Graph.* (2010), 1–10.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

Petr Savickỳ and Jaroslava Hlavácová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics* 9, 3 (2002), 215–231.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

Ana Serrano, Bin Chen, Chao Wang, Michal Piovarči, Hans-Peter Seidel, Piotr Didyk, and Karol Myszkowski. 2021. The effect of shape and illumination on material perception: model and applications. *ACM Trans. Graph.* 40, 4 (2021), 1–16.

Ana Serrano, Diego Gutierrez, Karol Myszkowski, Hans-Peter Seidel, and Belen Masia. 2016. An intuitive control space for material appearance. *ACM Trans. Graph.* 35, 6 (2016).

Weiqi Shi, Zeyu Wang, Cyril Soler, and Holly Rushmeier. 2021. A Low-Dimensional Perceptual Space for Intuitive BRDF Editing. *Eurographics Symposium on Rendering - DL-only Track* (2021), 1–13.

Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1633–1642.

Paul J Somerfield, K Robert Clarke, and Ray N Gorley. 2021. Analysis of similarities (ANOSIM) for 2-way layouts using a generalised ANOSIM statistic, with comparative notes on Permutational Multivariate Analysis of Variance (PERMANOVA). *Austral Ecology* 46, 6 (2021), 911–926.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* 33 (2020), 16857–16867.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *AAAI Conference on Artificial Intelligence* (2017).

Katherine R Storrs, Barton L Anderson, and Roland W Fleming. 2021. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour* 5, 10 (2021), 1402–1417.

Substance 3D Assets. 2023. https://substance3d.adobe.com/assets.

Tiancheng Sun, Ana Serrano, Diego Gutierrez, and Belen Masia. 2017. Attribute-preserving gamut mapping of measured BRDFs. *Computer Graphics Forum* 36, 4 (2017), 47–54.

William Thompson, Roland Fleming, Sarah Creem-Regehr, and Jeanine Kelly Stefanucci. 2011. *Visual perception from a computer graphics perspective.* CRC press.

Hiroyuki Tsuda, Munendo Fujimichi, Mikuho Yokoyama, and Jun Saiki. 2020. Material constancy in perception and working memory. *Journal of Vision* 20, 10 (2020), 10–10.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

Peter Vangorp, Jurgen Laurijssen, and Philip Dutré. 2007. The influence of shape on the perception of material reflectance. In *ACM Trans. Graph.* 77–es.

Gregory J. Ward. 1992. Measuring and Modeling Anisotropic Reflection. *Proceedings of ACM SIGGRAPH* 26, 2 (1992), 265–272.

David I Warton, Stephen T Wright, and Yi Wang. 2012. Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution* 3, 1 (2012), 89–101.

Josh Wills, Sameer Agarwal, David Kriegman, and Serge Belongie. 2009. Toward a perceptual space for gloss. *ACM Trans. Graph.* 28, 4 (2009), 1–15.

Chenyun Wu, Mikayla Timm, and Subhransu Maji. 2020. Describing textures using natural language. *European Conference on Computer Vision* (2020), 52–70.

Zelai Xu, Tan Yu, and Ping Li. 2022. Texture BERT for Cross-modal Texture Image Retrieval. *Proceedings of ACM International Conference on Information & Knowledge Management* (2022), 4610–4614.

Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2022. TileGen: Tileable, Controllable Material Generation and Capture. *Proceedings of ACM SIGGRAPH Asia* (2022).