Let's Go Shopping (LGS) – Web-Scale Image-Text Dataset for Visual Concept Understanding

Yatong Bai[1] Utsav Garg[2] Apaar Shanker[2] Haoming Zhang[2] Samyak Parajuli[2]
Erhan Bas[2] Isidora Filipovic[3] Amelia N. Chu[3] Eugenia D Fomitcheva[3] Elliot Branson[2]
Aerin Kim[2] Somayeh Sojoudi[1] Kyunghyun Cho[3]
[1]University of California, Berkeley [2]Scale AI [3]New York University
Correspondances to yatong_bai@berkeley.edu, aerinykim@gmail.com

Abstract

Vision and vision-language applications, such as image classification and captioning, rely on large-scale annotated datasets that require non-trivial data-collecting processes. This time-consuming endeavor hinders the emergence of large-scale datasets, limiting researchers and practitioners to a small number of choices. Therefore, we seek more efficient ways to collect and annotate images. Previous initiatives have gathered captions from HTML alt-texts and crawled social media postings, but these data sources suffer from noise, sparsity, or subjectivity. For this reason, we turn to commercial shopping websites whose data meet three criteria: cleanliness, informativeness, and fluency. We introduce the Let's Go Shopping (LGS) dataset – a large-scale public dataset with 15M image-caption pairs from publicly available e-commerce websites. When compared with existing general-domain datasets, the images of LGS have a clearer focus on the foreground object and fewer complex backgrounds. Our experiments on LGS show that the models trained on current benchmark datasets do not readily generalize to e-commerce data for the classification setting, while visual feature extractors can generalize in specific cases. Furthermore, LGS's high-quality e-commerce-focused images and bimodal nature make it advantageous for vision-language bi-modal tasks: LGS enables image-captioning models to generate richer captions and helps text-to-image generation models achieve e-commerce style transfer.

1. Introduction

Computer vision (CV) and natural language processing (NLP) tasks increasingly rely on pre-trained representations. While NLP representations can be trained on unannotated raw text, vision applications often consider pre-training using large-scale datasets with discrete class labels annotated by humans, such as ImageNet [8, 49] or OpenImages [28]. Vision-language bimodal applications, such as image captioning or visual question answering, similarly rely on large amounts of annotated data. Unfortunately, many of the large-scale bi-modal datasets now in existence, such as CLIP [46], ALIGN [23], and JFT300M [7, 18], are not publicly accessible. As a result, research has been constrained to a few selected large datasets, such as Conceptual Captions [4] and Microsoft COCO [5]. This shortage of available public datasets can be attributed in part to the time and effort required to gather, clean, and annotate massive datasets.

Therefore, we adopt a more efficient and scalable high-quality data collection pipeline to acquire image-text pairs that are easily available on e-commerce websites. While some existing datasets use public websites as annotation sources, most of them use social media websites (RedCaps [10]) or alt-texts[1] (Conceptual Captions [51]) for annotation sources. Nevertheless, social media data suffer from subjectivity. On the other hand, alt-texts can be excessively noisy, sometimes merely including uninformative texts such as "alt img", as shown in Figure 1.

As a result, we gravitate to e-commerce websites, where clean images with objective, accurate, succinct, and informative descriptions are abundant, as illustrated in Figure 2. Let's Go Shopping (LGS) dataset collects 15 million image-description pairs from approximately 10,000 e-commerce sites selling a wide range of products. Due to the nature of e-commerce data, the majority of LGS images have a clear background and a static focus on the stated object. On the captions front, LGS provides precise and elaborative captions. We show how highly precise information can be extracted from captions for vision-language fine-tuning.

On the other hand, ImageNet-1k has served as the ubiquitous go-to pre-training and evaluation dataset for vision-only applications. While ImageNet covers a wide range of domains, the diversity of angles and arrangements is restricted.

[1]Alt-texts are short descriptions of HTML website images. When an image cannot be rendered, the website displays its alt-text as a surrogate.

Figure 1. In comparison to e-commerce product descriptions, alt-text is usually less informative, sometimes too broad, or even irrelevant.

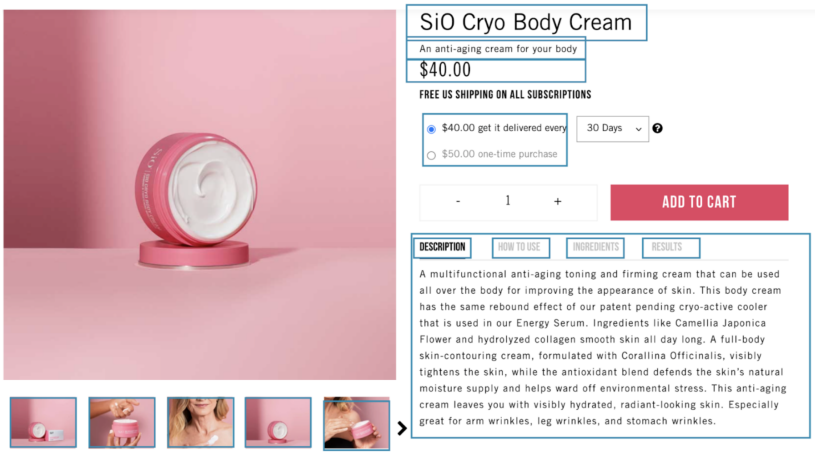


Figure 2. An e-commerce-based LGS sample instance with image, title and description.

As a result, the literature has shown that ImageNet models do not generalize well to deliberately-constructed out-of-distribution (OOD) scenarios [2]. This work use image classification experiments to demonstrate that such OOD data is ubiquitous in e-commerce applications. We then show that models can benefit from the unique e-commerce distribution in classification, reconstruction, captioning, and image generation tasks.

Specifically, we convert the LGS captions into taxonomies and labels and demonstrate a large disparity between the label distributions of LGS and ImageNet: even with best efforts, only 17.6% of the concepts are shared between popular ImageNet-1k synsets and the e-commerce corpus (more details in Section 3.4). Even for those shared classes, the performance of ImageNet models degrades significantly. By verifying that the LGS classes are well-separable, we conclude that this performance degradation can be mostly attributed to the distributional disparity. To separate the effects of labels and captions and isolate the distribution shift of the images, we consider Masked AutoEncoder (MAE) [16], a self-supervised pre-training method that does not rely on labels. We show that an MAE model trained on ImageNet-1k can reconstruct LGS images well, but adding LGS to the training data improves the performance on LGS and generalizes better to COCO.

The above results demonstrate that while the e-commerce images are from a distribution that is distinct from current benchmark datasets, the feature extractors can be shared. Moreover, we illustrate additional merits of LGS that qualify it as a pre-training dataset. Specifically, the models learned on both LGS and ImageNet have improved linear probing performance on common downstream tasks such as CIFAR-100 [27] and Fashion MNIST [60], compared with the ImageNet-only counterparts.

The distinctive distribution of LGS also benefits vision-language bimodal tasks. For caption generation tasks, we train an OFA model [58] on LGS to demonstrate that the clearer image foreground, cleaner image background, and the highly descriptive captions of LGS enable the model to produce "attribute-rich" image captions, which models trained on traditional datasets fail to produce.

For text-to-image generation tasks, we use Stable Diffusion (SD) [47] and fine-tune it in both general and fine-grained settings on subsets of the LGS dataset. We demonstrate promising qualitative and quantitative results on adapting existing text-to-image models using LGS for e-commerce-related generations. Furthermore, with the help of its distinct image style and descriptive captions, LGS can help the SD model generate e-commerce-styled images.

To make LGS available to the public, we will share the filtered links to the image-caption pairs, as was the case for ImageNet. We will also share the downloader so that the exact same dataset can be reproduced.

2. Related work

2.1. Unimodal Pre-Training Datasets

Prior to the popularization of bi-modal training, uni-modal data (vision-only or language-only) have been the workhorses for pre-training tasks. On the vision side, ImageNet-1k and ImageNet-22k are still some of the most prevalent examples, alongside the larger JFT-300M dataset. For the e-commerce domain, Fashion MNIST, Clothing1M [61], Fashion200k [15], and FashionIQ [59] have been proposed to analyze the effects of noisy labels. For general wide-domain downstream tasks, CIFAR-10, CIFAR-100, MNIST [30], SVHN [42], and Tiny ImageNet [29] are some of the most common choices.

Datasets	Instances	Released
Let's Go Shopping (this paper)	14,847,764	✓
YFCC100M (Yahoo)	100 million	✓
RedCaps (University of Michigan)	12,011,111	✓
Conceptual Captions 12M (Google)	12,423,374	✓
WIT-English (Google)	5,500,746	✓
Localized Narratives (Google)	849,000	✓
COCO (Microsoft)	328,000	✓
Visual Genome (Stanford)	108,077	✓
CLIP (OpenAI)	400M	×
ALIGN (Google)	1.8B	×

Table 1. The instance count of LGS compared with existing bi-modal datasets.

2.2. Vision-and-Language Pre-Training Datasets

The literature has shown that image-text data from COCO can be used to learn *visual* features that are competitive with supervised pre-training [17] on ImageNet when transferred to downstream tasks [3, 9, 12, 14, 36, 57, 64]. More recently, CLIP and ALIGN scaled up to 400M and 1B+ web-curated image-text pairs, enabling zero-shot visual recognition on downstream tasks.

Originally intended for image-text retrieval and image captioning, bi-modal datasets are now widely used for training cross-modal representations [6, 20, 25, 32, 33, 35, 38, 43, 51, 53, 55, 65] that transfer to downstream tasks, such as visual question answering [1, 21, 66], referring expressions [24], and visual reasoning [54, 63]. In light of these novel training paradigms, more recent works build larger datasets specifically for vision-and-language pre-training. Examples include LAIT [45], Conceptual Captions-12M, and Wikipedia-ImageText (WIT) [52], Localized Narratives [44], Visual Genome [26], YFCC100M [56]. Similar to these datasets, LGS offers rich semantic data for pre-training applications. However, our choice of the e-commerce data source is unique, leading toward distinctive data distribution.

Image-text datasets are also used for learning visual features. The work [31] has proposed to train visual n-gram models on YFCC100M, whereas other methods [3, 9] aim to learn features from the captions from the COCO dataset [5]. The quality of the resulted features are competitive with supervised ImageNet training [17] on many downstream tasks [12, 14, 36, 49, 57]. Moreover, the image-text pre-training schemes scale up to very larger non-public datasets that are even larger than LGS [23, 46].

A core motivation for collecting image-text pairs from the internet is the possibility of scaling up the data size without bearing the prohibitively expensive annotation costs. In light of this motivation, there have been multiple efforts of collecting large quantities of noisy labels associated with online images, leading to datasets such as WebVision [34], YFCC100M, JFT-300M, and Instagram-3.5B [39].

Existing multi-modal e-commerce-inspired datasets include M5Product [11] and DeepFashion [37]. With 6 million instances, M5Product's size is around a half of LGS's. While M5Product focuses on demonstrating the effectiveness of multi-modal training, this paper emphasizes analyzing the e-commerce data distribution and how it generalizes to general wide-domain datasets in a pre-training setting.

3. The Let's Go Shopping (LGS) Dataset

With 14,847,764 image-text pairs, the LGS dataset has a size advantage over many publicly available bi-model datasets, as presented in Table 1. In this section, we offer additional analysis of the LGS data. For all analysis and experiments in the paper, we use a subset of the instances with 13 million instances, as the rest of the dataset was constructed in parallel with the experiments.

3.1. Data Collection

To create training data that is truly representative of e-commerce data as a whole, we include a wide range of commerce websites with various product kinds, such as infant products, sporting goods, bridal jewelry, etc.

The collection pipeline starts with a set of heuristic rules to isolate the product pages from the non-product pages of an e-commerce website. Then, our automated extractor obtains relevant information on each product page, including the product title, the description, and the first listed image. Some products may include numerous variants (e. g., different colors for a type of T-shirt), and we collect all variants. We avoid crawling information that the sellers are unwilling to share. Specifically, the extractor is forbidden from crawling pages with a 'Disallow' extension. Finally, we use strict automated tests to filter out the instances with potential quality issues. Examples of the tests include confirming that the price is a number, certifying that the images are valid, and ensuring that the product title exists and contains no unexpected characters.

3.2. Characteristics of LGS Images

In general-domain image-caption datasets, the images usually consist of one or more subjects juxtaposed against a rich background, and their captions often use the background as the context. In contrast, e-commerce product thumbnails in LGS often depict only one in-animate item that occupies the foreground without any association with the background. The background is also often a single color, with some examples shown in Fig. 3. These clear backgrounds make it easier for models to locate the patterns that correspond to their tasks.

3.3. Characteristics of LGS Captions

In this subsection, we analyze the traits of the LGS captions. The LGS dataset has 14,847,764 captions in total and

Dataset	Min	Max	Mean	Var	Median	Top-5 most frequent	Skew
LGS	2	3642	89.58	6471.92	67	[34, 30, 47, 31, 49]	3.44
COCO	5	50	10.56	6.25	10	[9, 10, 8, 11, 12]	2.76

Table 2. Comparing the word count statistics of the LGS and COCO captions.

Figure 3. Examples of LGS images with taxonomy end leaves

Gender	Count	Percentage
Women's	6,039,577	31.2 %
Men's	2,347,228	14.5 %
Unisex	1,767,303	10.9 %
Unknown	6,072,933	31.4 %

Table 3. The instance count of LGS by gender.

Dataset	C. Nouns	P. Nouns	Adjectives	Verbs
LGS	158,479	139,174	48,907	57,481
COCO	10,403	1,655	3,053	4,961

Table 4. The POS's that occur at least ten times.

the words and phrases in LGS captions are diverse. For example, while LGS has around 3x more captions than COCO[2], its captions possess about 20x more uni-grams, bi-grams, and tri-grams, with more detailed statistics presented in Appendix A.5. Table 2 presents some statistics of the word distribution of the captions, showing that both LGS and COCO have highly positively skewed distributions, with LGS having a longer tail. Since LGS incorporates data from a large variety of e-commerce websites, the descriptions can include rich information. In the subsequent sections, we show that while the raw captions of LGS are diverse, clear structural information can be extracted from the LGS captions for fine-tuning purposes.

Additionally, we use the part-of-speech (POS) tagging method from the Spacy library [19] to analyze the linguistic statistics of the LGS captions, comparing common nouns, proper nouns, adjectives, and verbs. Table 4 illustrates that LGS has at least 10x more words per POS compared with COCO, whereas Figures Supp-5 and Supp-6 in the supplementary materials provide further insights into the composition of each word type. Due to the e-commerce nature of LGS, a large portion of the instances is clothing and other wearable items. Thus, within LGS, the proper nouns often present the brand names and sizes, the common nouns often describe the materials, and the adjectives and verbs often characterize the product-specific descriptions and actions, making the LGS captions highly descriptive.

The numbers of gender occurrences among all LGS instances are presented in Table 3.

3.4. LGS for Classification

While the raw data format of LGS is image-caption pairs, we also experimented with image classification with LGS by labeling the classes. Specifically, we build three classification variants: LGS-117, LGS-710, and LGS-Overlap. For all three variants, we use a taxonomy generation language model pre-trained in-house to convert each product description into a taxonomy tree, whose nodes are designed to be informative for e-commerce catalog applications. The end leaf of each taxonomy tree is then used as the label, with some examples displayed in Figure 3. The taxonomy tree can also be used to generate summarized image captions that include *product title*, *product brand name*, and a number of *"bullet strings"* describing specific product attributes. The bullet strings include examples such as `Nylon fabric`, `Classic collar`, and `Front zipper fastening`. The LGS leaves form a long-tailed distribution that emphasizes common daily commodities, with the five most common leaves being `Tops and T-shirts`, `Dresses`, `Rings`, `T-shirts`, and `Sweatshirts and Hoodies`. For each of the three classification variants, we further clean the end leaves, with

[2]Each COCO instance has five corresponding captions, and we consider each of them separately.

details provided in the two following paragraphs. In Figure Supp-1 in the supplementary materials, we provide a histogram of the end leaf distribution.

LGS-117 and LGS-710 are designed as pre-training datasets. Within all raw labels generated by the taxonomy model, there are synonyms and overlaps that should be unified. After manually merging the synonyms among the most popular classes, we observe 117 classes that contain at least 10k images. We select 10k images from each class, forming the balanced LGS-117 dataset. LGS-710 is an unbalanced dataset that includes more scarce classes. To accelerate label engineering, we use a semi-automated pipeline. First, we remove uninformative words like "other" and parse juxtaposed nouns by commas and "and". Next, we use a pre-trained language model to extract the embedding of each parsed noun. As an example, for the leaf `Tops and T-shirts`, we embed both `tops` and `t-shirts`. We then consider the "similarity" between two classes to be the maximum cosine similarity between all pairs corresponding nouns. Very close classes are merged based on a similarity threshold of 0.92, which is determined by manually inspecting the merged classes.

LGS-Overlap is proposed as an out-of-distribution test set for models trained on ImageNet-1k, one of the most widely-used benchmarking datasets. We use a similar semi-automated pipeline to merge LGS classes with ImageNet synsets [8, 41]. We optimize the pipeline by adjusting the similarity threshold to 0.90 and including additional preprocessing steps such as singularization and keyword merging. Note that polysemous words in the labels can refer to different objects in LGS and ImageNet. For example, "cricket" in LGS refers to sports equipment but refers to the insect species in ImageNet. Thus, a manual inspection of the merged classes is performed. After discarding classes with less than 20 instances, we gather the remaining 176 ImageNet synsets that align with the LGS end leaves and use them as the LGS-Overlap dataset. The fact that only 17.6% of the ImageNet synsets are matched shows a significant label distribution difference between e-commerce applications and common pre-training datasets. Since a higher level of label-space alignment is essential for more effective pre-training [39], LGS forms a representative benchmark and a pre-training dataset for downstream tasks that see distributions close to e-commerce.

4. Experiments

4.1. Image classification and reconstruction

In this subsection, we use image classification and reconstruction tasks to characterize the distributional difference between LGS and ImageNet. We consider the distributions of images as well as the labels.

4.1.1 ImageNet models do not readily generalize to E-commerce

The existing literature has shown that carefully-constructed images collected in a bias-controlled manner can elicit a significant performance degradation on classifiers trained on ImageNet [2]. By applying pre-trained ImageNet classification models to the LGS-Overlap dataset without further training, we show that such out-of-distribution examples naturally exist in the e-commerce domain. Specifically, we use publicly-available weights of a ResNet-50 model and a ConvNeXT-Base model. The ResNet-50 achieves a 74% average recall across the 176 overlapping synsets over the ImageNet images, but the number noticeably reduces to 46.43% on LGS-Overlap. The ConvNeXT-Base obtains 79.00% and 50.14% on ImageNet and LGS-Overlap, respectively. This difference highlights that existing ImageNet models do not readily transfer to LGS instances. In addition to having a different label distribution, the e-commerce domain forms a natural distribution shift even for the classes that also exist in ImageNet. While taxonomy standardization techniques exist, aligning and merging the label space is still hard in general. Thus, a pre-training dataset that is more aligned with e-commerce is necessary, and LGS fulfills this role.

We further show that LGS end leaves are well-separable, verifying that the performance degradation of ImageNet models is caused by the distribution mismatch and not the ambiguity of the LGS classes. Note that Table 5 illustrates that the models learned on LGS-117 / LGS-710 can achieve high accuracy on LGS-117 / LGS-710. Specifically, we consider the "linear probing followed by fine-tuning" training schedule, a transfer learning scheme that has been shown to improve the robustness against distribution shift by avoiding significant distortions of the pre-trained weights.

4.1.2 Visual feature extractors can generalize

Since the image-label correspondence is different between LGS and ImageNet, we use self-supervised training to isolate this mismatch and focus on the distribution of images. In the context of transfer learning, since self-supervised training does not use labels, it circumvents the issue of label space mismatch between target and source domains, which has been shown to undermine the quality of transfer learning. Masked AutoEncoder (MAE) [16] is a self-supervised method designed for pre-training. Thus, we compare the performance of an MAE trained on ImageNet only with an MAE trained on ImageNet and LGS-710. Figure 4 shows that the MAE trained on ImageNet can reconstruct a reasonable LGS image, but the reconstruction quality of the ImageNet+LGS model is better, demonstrating that LGS can be used to learn e-commerce visual features.

To quantitatively demonstrate the generalizability of the vision feature extractors, we evaluate the reconstruction per-

LGS Accuracy	LGS-117 from scratch	LGS-117 IN-pretrained	LGS-710 IN-pretrained (Top-1)	LGS-710 IN-pretrained (Top-5)
After linear probing	–	69.58 %	60.72 %	81.16 %
After fine-tuning	97.89 %	98.16 %	77.27 %	89.09 %

Table 5. The classification accuracy of models trained on LGS shows that the LGS end leaves are well-separable.

Training Dataset	Inception (↑)	FID (↓)
ImageNet-1k	9.2930	114.60
IN pretrain→IN+LGS	9.1906	115.48
LGS	10.187	91.387

Table 6. The reconstruction quality of the MAE models trained on LGS and ImageNet, evaluated on COCO. The symbol ↑ denotes "higher is better" while ↓ means "lower is better".

Figure 4. While an MAE trained on ImageNet can reasonably reconstruct an LGS image, adding LGS instances to the training improves the reconstruction quality.

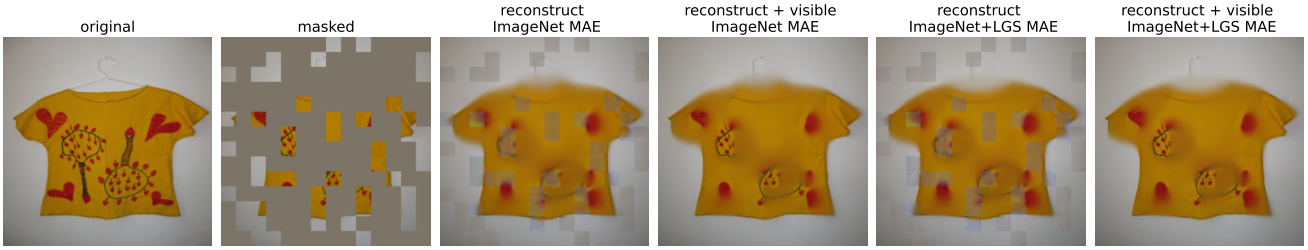


Figure 5. Adding LGS instances to the training also improves the reconstruction on some ImageNet instances.

Linear probing dataset	MAE Training Setting A	B	C
LGS-117 (40 epochs)	72.98 %	76.37 %	76.87 %
ImageNet-1k (60 epochs)	67.78 %	46.37 %	65.29 %

Table 7. Linear probing accuracy of the self-supervised MAE models initialized by three different initializations. A: baseline imagenet MAE model [16], B: LGS MAE model, C: LGS+Imagenet MAE model. A and B are cold-start models, whereas C is initialized with A followed up by 150 epochs on mixed Imagenet and LGS-710 data (ratio 1:1). In finetuning LGS-117 and Imagenet datasets we used 40 and 60 epochs respectively.

formance of the MAE models trained on LGS and ImageNet on COCO. The qualities of the raw reconstructions obtained by the models are presented in Table 6. While LGS is more domain-specific compared with ImageNet and COCO (both of which cover a wide range of domains), the MAE trained on LGS is able to generate COCO images with higher qualities compared with the ImageNet model. Furthermore, we use Table 7 to show that upon the visual embeddings learned jointly on ImageNet and LGS, a linear classifier with a satisfactory performance can be learned on both ImageNet and LGS. The above results verify that the feature extractors can generalize between LGS and general-domain datasets, despite the separation of the intermediate visual embeddings (which are visualized in Appendix A.2).

Based on the above observations, we infer that the e-commerce data distribution, represented by the LGS dataset, significantly differs from existing general datasets in the label space, while visual features can generalize. Thus, LGS is an ideal pre-training dataset for downstream tasks whose class distributions align with the e-commerce domain.

4.1.3 LGS supplements ImageNet as a pre-training dataset

LGS can also widen the span of the pre-training distribution when used in conjunction with ImageNet, acting as a bridge between general visual features and domain-specific applications. Specifically, Table 8 shows that a two-phase ImageNet→LGS-710 weakly-supervised pre-training scheme produces features more suitable for fine-tuning on

Pre-training Setup	Linear Probing					End-to-end training	
	CIFAR-10	CIFAR-100	Fashion MNIST	Clothing1M (10 %)	Clothing1M (100 %)	Clothing1M (10 %)	Clothing1M (100 %)
ImageNet	61.97	40.46	79.68	59.74	67.57	65.69	74.81
ImageNet→LGS-117	59.83	35.57	80.39	64.48	69.67	68.16	75.47
ImageNet→LGS-710	58.81	42.21	82.18	64.16	70.06	65.85	74.51

Table 8. ImageNet→LGS-710 two-phase pre-training improves downstream linear probing accuracy for downstream tasks including CIFAR-100, Fashion MNIST, and Clothing1M. On Clothing1M, whose data also comes from the e-commerce domain, the LGS-pre-trained features also improve end-to-end fine-tuning performance. For Clothing1M, we only use its clean training set, whereas Clothing1M (10%) is a few-shot setup that trains on a 10% subset of the clean training set.

common downstream tasks. On e-commerce-related downstream datasets such as Clothing1M, the models pre-trained on LGS also excel in both linear probing and end-to-end settings.

In linear probing experiments, we observe that incorporating in-domain pre-training (both LGS-117 and LGS-710) results in better performance (2% absolute) compared to ImageNet pre-training. Moreover, in limited-data settings, we observe less model regression compared to the full-data setups. For example, for fine-tuning a linear classifier on 10% of the Clothing1M-clean dataset, the ImageNet pre-trained model regresses more (11.5% relative) compared to LGS-117 and LGS-710 pre-trained models (7.4 and 8.4% relative respectively). When models are trained end-to-end, we observe that the pre-training setup is less critical on fine-tuning the full Clothing1M-clean training dataset. However, for limited-data experiments, filtering out under-represented classes (LGS-117) in pre-training helps with the downstream fine-tuning results (2% absolute) compared to both ImageNet and LGS-710 datasets.

In Appendix A.3 in the supplementary materials, we use GradCam [13, 50] to visualize the representations learned by the classification models, demonstrating that the LGS models look for much more localized patterns that are relevant to e-commerce classification.

4.2. Caption Generation using OFA

In this section, we illustrate that the distinct distribution of LGS benefits vision-language bi-modal tasks. Specifically, we study the efficacy of image-captioning (IC) models trained on traditional datasets in predicting LGS-type descriptions. We also evaluate the performance of LGS-trained models in generating attribute-rich image captions that would otherwise not be possible for models trained on more traditional datasets.

In this experiment, we utilize a bi-modal modeling framework based on OFA [58], a recently-proposed encoder-decoder architecture that has achieved state-of-the-art performances in many language-vision tasks. For each LGS image, the corresponding caption can be constructed by concatenating the "product description" strings in various orders. Specifically, we create three types of captions:

1. LGS-title : title and brand name;
2. LGS-taxonomy : product taxonomy;
3. LGS-description: concatenated bullet strings.

Training Set	Test Set	METEOR (↑)
LGS-title	LGS-title	0.184
LGS-description	LGS-title	0.161
LGS-taxonomy	LGS-taxonomy	0.584
COCO	LGS-title	0.069

Table 9. IC model performance of image-captioning task evaluated on different combinations of training and evaluation datasets.

The OFA IC model was trained on the three types of LGS inputs as well as on the traditional COCO dataset. The IC model performance in terms of its ability to predict the appropriate target string is tabulated in Table 9.

4.3. Text-to-Image Generation

Because of its high-quality e-commerce-focused images and bimodal nature, LGS is an ideal option for training text-to-image models in the e-commerce sector, serving as a bridge between general visual features and domain-specific applications. In this section, we use LGS to adapt the Stable Diffusion (SD) text-to-image generation method to two e-commerce scenarios: general and fine-grained. For both scenarios, we fine-tune based on the `sd-v1-4` (referred to as Vanilla) checkpoint of SD.

For the general setting, we add a domain identifier to all training prompts associated with LGS images and guide the SD model to adapt to the e-commerce image style when this identifier is provided. The choice of the domain identifier is crucial, as the paper [48] shows that a domain identifier with a strong prior should be avoided. For example, the word `retail` has a strong prior, and the pre-trained "Vanilla" SD model confidently associates it with (physical) retail stores.

Figure 6. Qualitative comparisons of the generations of the Vanilla and the LGS-117-fine-tuned SD models in the general setting. The fine-tuned model generates more visually appealing images.

This behavior is undesirable for the goal of e-commerce style transfer. By analyzing the effects of various domain identifiers on the generations of the pre-trained SD model, we determine that the word "e-commerce" gives a weak prior and is a suitable identifier. We then construct the ground-truth training prompts for the LGS images in the format of `a photo of <brand> <end_leaf> <title>, e-commerce`, where the `<end_leaf>` refers to the end leaf of the taxonomy tree introduced in Section 3.4. The "Vanilla" SD checkpoint is fine-tuned on one million LGS image-prompt pairs for 100k steps with a batch size of 24. Table 10 displays the quantitative results on an unseen validation set (5K image-prompt pairs) from LGS and a subset of the DeepFashion InShop dataset. The fine-tuning process enhances the performance of SD on LGS as expected. While the FID scores on DeepFashion are lower, the generations of the LGS-117 fine-tuned model are aesthetically more appealing. At this instant, there are no quantitative metrics that directly measure aesthetic superiority. Thus, we present Figure 6 and the additional examples in Appendix A.6 in the supplementary materials (Figures Supp-9 and Supp-8) to demonstrate the aesthetic improvement qualitatively. The

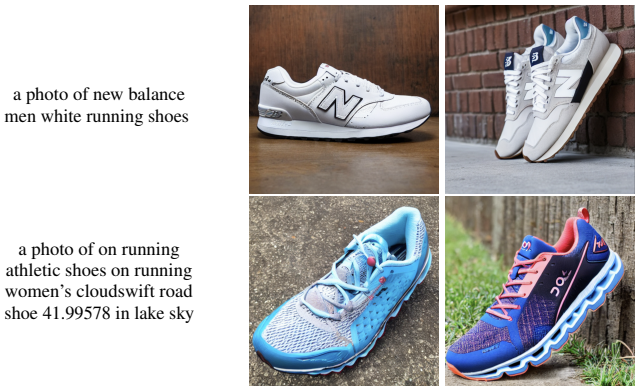


Figure 7. The LGS-117-fine-tuned SD model also generates more visually appealing images in the fine-grained setting. The prompts are from LGS.

Model	Test Set	FID (↓)
Vanilla	LGS Val	25.3498
Vanilla + LGS-117	LGS Val	**24.1952**
Vanilla	DeepFashion	**62.9269**
Vanilla + LGS-117	DeepFashion	74.0185

Table 10. Comparing the Vanilla SD and the LGS-117 fine-tuned model on LGS and DeepFashion datasets.

lower FID scores may indicate a distribution shift between LGS and DeepFashion images.

For the fine-grained setting, we use data belonging to only a particular end leaf, using the same prompt without the additional identifier. The checkpoint is fine-tuned with 10k image-prompt pairs for 25k steps with a batch size of 6. We use the "athletic shoes" end leaf as an example and compare the generations before and after LGS-fine-tuning under the fine-grained setting in Figure 7. As did the general setting results, the fine-grained examples also indicate that LGS helps adapt text-to-image models to e-commerce scenarios and improves image quality and aesthetics.

5. Conclusion

The Let's Go Shopping (LGS) dataset consists of 15 million pairs of publically-accessible diverse images and descriptive captions from e-commerce websites. Our efficient semi-automated gathering and annotation pipeline ensure scalable data collection. We then use LGS to show that while the categories associated with e-commerce data may not align with the general-domain pre-training datasets, visual feature extractors can be shared. Finally, we show that the distinct distribution offered by LGS and LGS's bi-modal nature can be beneficial for applications including image classification, image reconstruction, bi-modal representation learning, and text-to-image generation.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 3

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 2019. 2, 5

[3] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *ECCV*, 2020. 3

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*, 2021. 1

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 3

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 3

[7] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR*, 2017. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1, 5

[9] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2020. 3

[10] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: web-curated image-text data created by the people, for the people. *ArXiv*, abs/2111.11431, 2021. 1

[11] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *CVPR*, 2022. 3

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2009. 3

[13] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 7, 12

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 3

[15] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, 2017. 2

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. 1

[19] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. 4

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *arXiv preprint arXiv:2004.00849*, 2020. 3

[21] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-world Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019. 3

[22] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 2019. 12

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv preprint arXiv:2102.05918*, 2021. 1, 3

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 3

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*, 2021. 3

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision using Crowdsourced Dense Image Annotations. *IJCV*, 2017. 3

[27] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2012. 2

[28] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[29] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. 2

[30] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs*, 2, 2010. 2

[31] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *ICCV*, 2017. 3

[32] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *AAAI*, 2020. 3

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[34] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv preprint arXiv:1708.02862*, 2017. 3

[35] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 3

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[37] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 3

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3

[39] Dhruv Kumar Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 3, 5

[40] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. 12

[41] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 5

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 2

[43] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NeurIPS*, 2011. 3

[44] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 3

[45] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv preprint arXiv:2001.07966*, 2020. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 3

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 7

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1, 3

[50] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 2017. 7, 12

[51] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning. In *ACL*, 2018. 1, 3

[52] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. *arXiv preprint arXiv:2103.01913*, 2021. 3

[53] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 3

[54] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019. 3

[55] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 3

[56] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 2016. 3

[57] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 3

[58] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 2, 7

[59] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021. 2

[60] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 2

[61] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. 2

[62] Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Peter Vajda, Kurt Keutzer,

and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with 2d image pretrained models. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, 2022. 13

[63] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 3

[64] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014. 3

[65] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. *AAAI*, 2020. 3

[66] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *CVPR*, 2016. 3

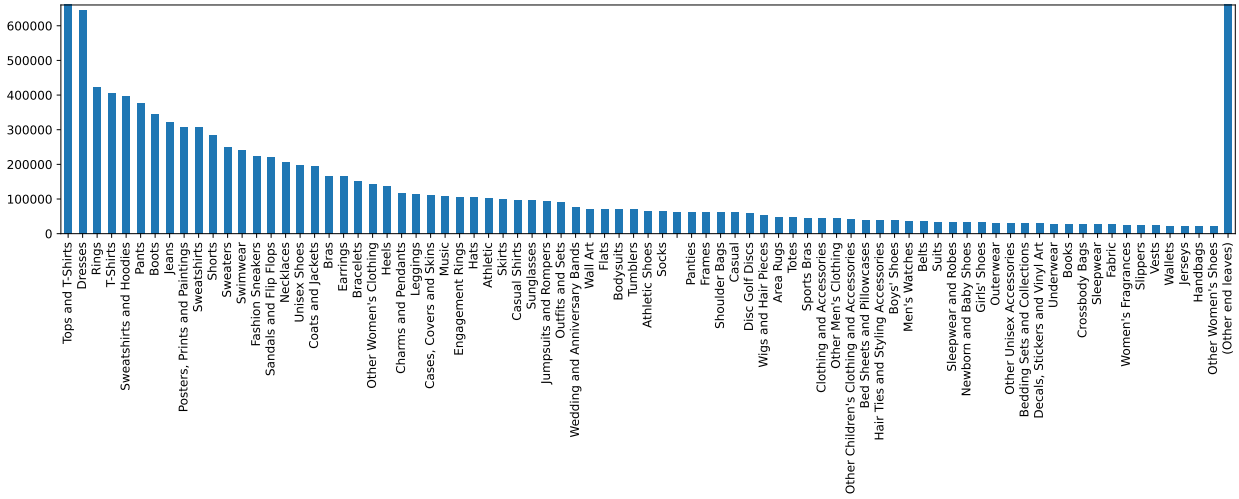


Figure Supp-1. The instance counts of the 80 most popular LGS end leaves.

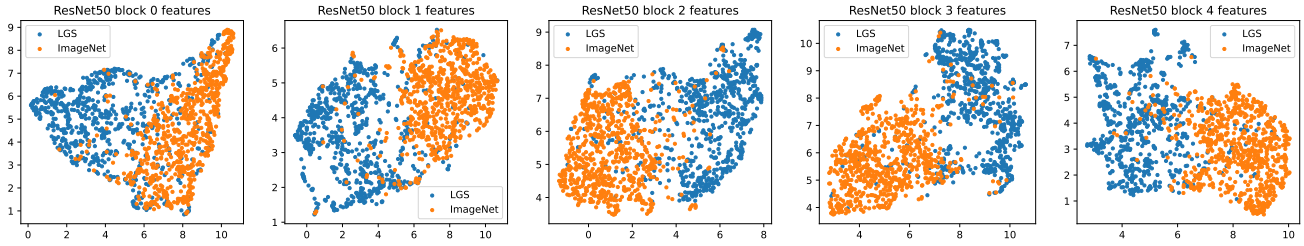


Figure Supp-2. UMAP visualization of the ImageNet and LGS features extracted on a ResNet50 model trained on ImageNet and LGS.

A. Additional analyses

A.1. LGS end leaf histogram

The instance counts of the LGS end leaves are displayed in Figure Supp-1. The top 80 most popular end leaves encompass 83.28% of the total instances, with the most popular `Tops and T-shirts` containing 16.23% of the instances.

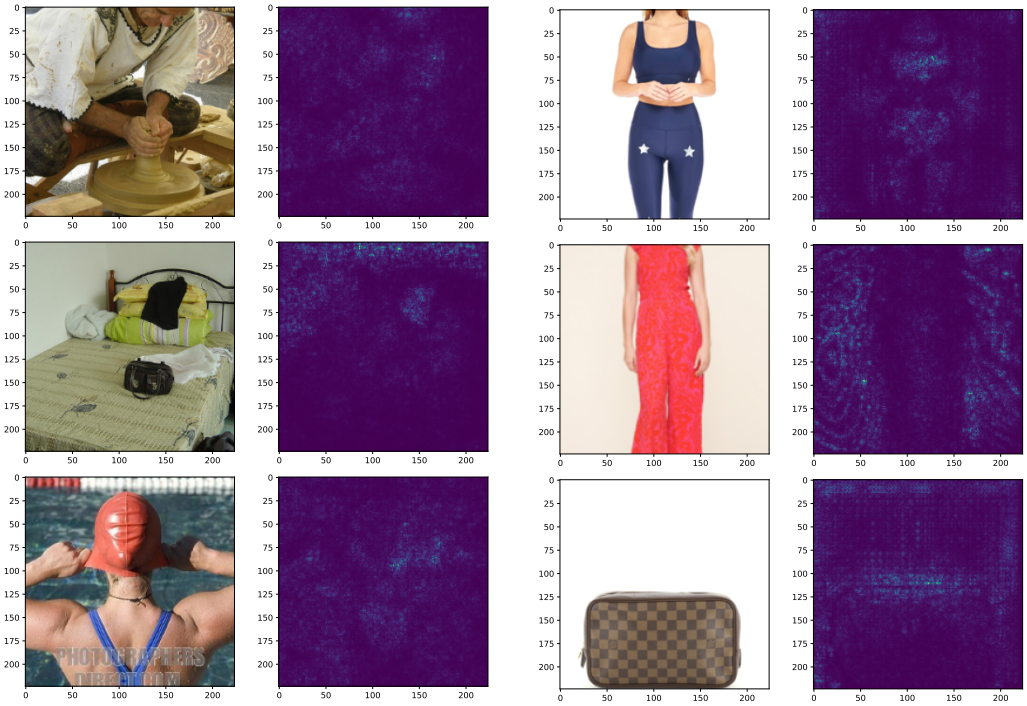
A.2. How do IN and LGS features differ?

To understand how vision models interpret the ImageNet and LGS instances, we use a ResNet50 model sequentially trained on ImageNet and LGS-117 as the feature extractor, and use UMAP [40] to visualize the high-dimensional ImageNet and LGS features in 2D figures. As shown in Figure Supp-2, the ImageNet features form a cluster, while the LGS features form a less concentrated cluster. The separation of the two clusters is especially prominent at the first two layers.

As discussed in the main portion of the paper, many LGS product thumbnails consist of isolated foreground objects and clear backgrounds, while ImageNet instances are mostly natural images where the foreground blends into the background. Thus, we question whether the feature clustering is a consequence of this difference. To this end, we learn a binary classification linear header that predicts between LGS and ImageNet images based on the features extracted by the ResNet-50 model. We then visualize the saliency map of this binary model in Figure Supp-3. While the background is the most prominent difference between ImageNet and LGS to human eyes, the saliency maps demonstrate that the deep models look for more sophisticated patterns, which can vary across different images. Specifically, the foreground is emphasized in the first LGS example, while the background is more important in the second LGS instance. This observation aligns with the findings of [22], which states that deep neural networks are not always understandable by humans.

A.3. LGS classification models look for localized patterns

In Figure Supp-4, we use GradCam [13, 50], a framework that visualizes gradient activations of the input images, to demonstrate that the models trained on LGS look for much more localized patterns. Here, we draw examples from the



(a) ImageNet examples. (b) LGS examples.

Figure Supp-3. The saliency map of the LGS-ImageNet binary classifier.

"sweatshirt" synset in the LGS-Overlap dataset, and feed them into the three ResNet-50 models learned on ImageNet, LGS-117, and LGS-710, respectively. The gradient activation of the ImageNet model spreads across the entire image, while the LGS models return more concentrated gradient maps. Note that the gradient spikes produced by the LGS models mostly locate around the sleeves and the waist portion of the clothes. This makes sense because the LGS models are trained to differentiate various kinds of clothing-dominated e-commercial products. The portions highlighted by the LGS model gradient maps precisely correspond to the places where various types of clothes differ. For example, checking the sleeve length may be one of the easiest ways of distinguishing T-shirts from sweatshirts. Since the LGS-710 model was trained to classify more fine-grained types of products, it looks for even more localized patterns compared with the LGS-117 model.

A.4. Linear probing setting details

In this section of the appendix, we discuss the implementation details for the linear probing experiments in Section 4.1.3. In the existing literature, when ResNets (designed for 224×224 inputs) are adopted for tasks that use smaller input sizes, the first 7×7 convolution layer is often replaced with a 3×3 layer. We adopt this replacement for CIFAR and Fashion MNIST. During linear probing, we thus allow this modified, randomly-reinitialized first layer to be optimized along with the output layer.

In Section 4.1.3, we presented the improved linear probing results on CIFAR-100 and Fashion MNIST. We would like to highlight that linear probing is a practical training method, because when the batch normalization (BN) layers are jointly optimized alongside the first and the last layer, this modified "linear" probing scheme can achieve a performance that is comparable to end-to-end training [62]. Specifically, with learnable BN, a ResNet-50 model pre-trained on ImageNet→LGS-710→ImageNet achieves an accuracy of 71.41% on CIFAR-100, compared with 69.47% for an ImageNet-only model.

A.5. n-gram and POS analysis of LGS captions

Table Supp-1 presents the comparisons of the uni-grams, bi-grams, and tri-grams of LGS. This comparison indicates that LGS is more linguistically diverse. The uni-grams and bi-grams of the two datasets are similar. However, we notice greater conceptual diversity for LGS within its tri-grams. Specifically, COCO's five most frequent tri-grams describe a group of objects and the relative position of the objects, whereas the LGS tri-grams encompass inherent properties of the commodities, including the size and the nature of each item.

Figure Supp-4. GradCam visualizations show that LGS classification models look for much more localized patterns.

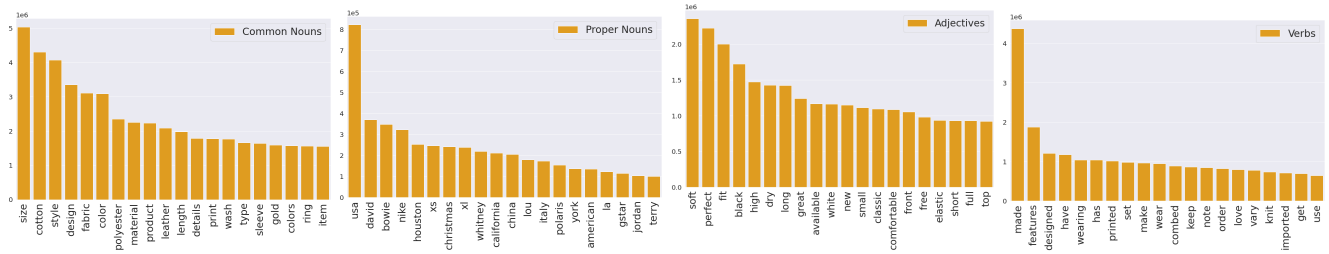


Figure Supp-5. Top 20 most common words per POS for LGS.

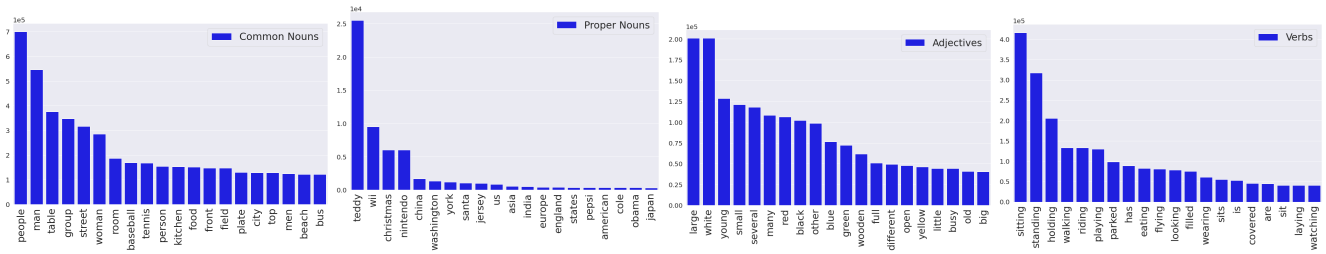


Figure Supp-6. Top 20 most common words per POS for COCO.

	Number of n-grams with occurrence ≥ 10		Five most frequent n-grams ($n = 1, 2, 3$)	
	LGS	**COCO**	**LGS**	**COCO**
uni-grams	364,802	17,009	and, the, a, to, with	a, of, on, the, i
bi-grams	4,054,418	184,882	with a, in the, of the, is a, for a	on a, in a, a man, of a, with a
tri-grams	8,900,084	462,653	true to size, made to order, this is a, this item is, machine wash cold	a group of, group of people in front of, next to a , on top of

Table Supp-1. Comparing the n-gram statistics of LGS with COCO.

Prompt ID	**Model**	**FID (↓)**	
		LGS	DeepFashion
1	Vanilla	**40.4437**	**61.8519**
	Vanilla + LGS-117	42.7328	74.4327
2	Vanilla	42.1081	**63.2344**
	Vanilla + LGS-117	**42.0529**	77.7190
3	Vanilla	36.7157	**58.2189**
	Vanilla + LGS-117	**36.1946**	79.3607
4	Vanilla	38.4101	**62.9269**
	Vanilla + LGS-117	**38.4100**	74.0185

Table Supp-2. Frechet Inception Distance (FID) results across prompts using a subset ($n = 5000$) of the LGS and DeepFashion InShop datasets.

In addition to the part-of-speech (POS) results presented in Section 3.3, we use Figures Supp-5 and Supp-6 to present the most common words per POS for LGS and COCO, respectively.

A.6. Determining the prompts for text-to-image generation

Ensuring the quality of the input prompts is paramount for text-to-image models to generate realistic images. Our goal is to choose a prompt which generates images faithful to the metadata, performs relatively well in terms of Frechet Inception Distance (FID) score, and generalizes across datasets.

To that end, we randomly selected 5,000 examples each from the LGS and DeepFashion InShop datasets. It is important to note that, for prompt engineering, the ground-truth images used for FID calculation are upscaled from 256×256, and the denoising diffusion implicit model steps (`ddim_steps`) were lowered to 50 for inference. This resulted in lower scores than the experiment results (Table 10). However, the numbers are still indicative of relative performance.

Quantitatively, Prompts 3 and 4 perform significantly better on LGS, perform comparably on DeepFashion, and generalize well (Table Supp-2). Prompt 3 had better FID scores using the Vanilla model and performed slightly better on LGS.

Qualitatively, however, Prompt 4 generations are consistently better and more faithful to the metadata (Figure Supp-7). Therefore, we choose Prompt 4 for our experiments. This also reaffirms that these metrics are not strong indicators of aesthetic quality in this particular case, and should only be used as a loose relative measure. Figures Supp-9 and Supp-8 show additional examples from the two datasets generated with Prompt 4.

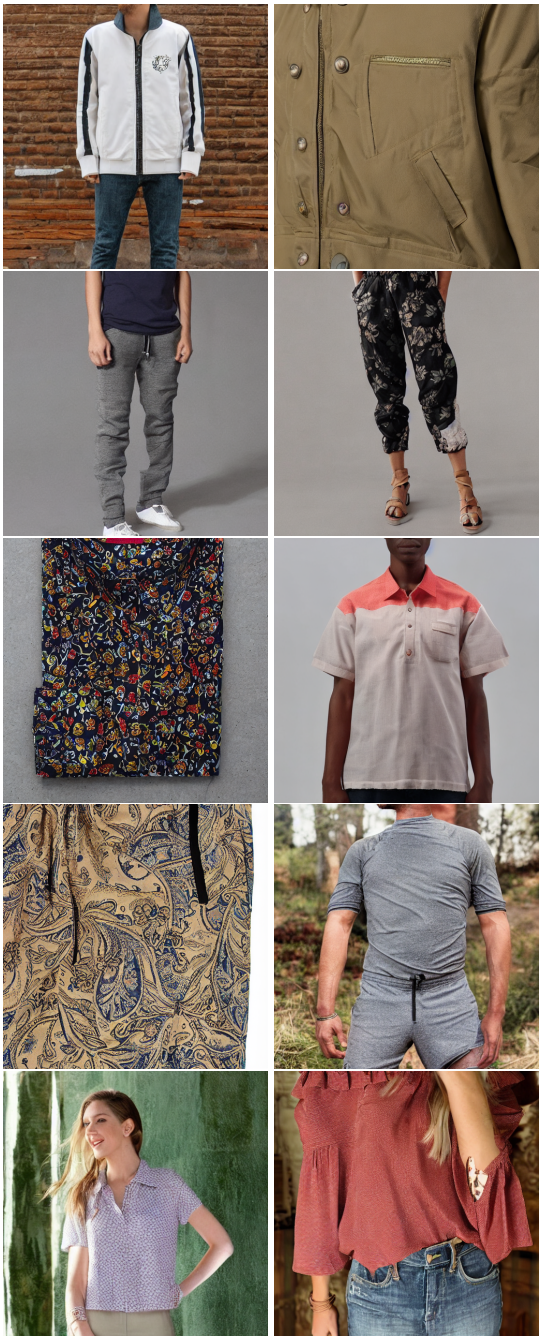
end_leaf: Jackets Vests
gender_category: Men
color: Khaki
first sentence of description:Made in a cotton-nylon blend with a modified collar and partial mesh lining, this baseball jacket is the slickest iteration of the style yet

end_leaf: Pants
gender_category: Men
color: Black-grey
first sentence of description: This jogger's easy, slouchy silhouette gets a little grit courtesy of its eye-popping print of photorealistic roses

end_leaf: Shirts Polos
gender_category: Men
color: Coral
first sentence of description: Constructed from cotton for a classic fit, this lightweight shirt features buttoned chest pockets

end_leaf: Shorts
gender_category: Men
color: Grey
first sentence of description: Crafted from speckled French terry, this sharper-than -average pair of sweatshorts is outfitted with a mock fly and three shiny zip pockets (two in front, one in back), ideal for lounging around or winning triathalons (just kidding)

end_leaf: Blouses Shirts
gender_category: Women
color: Rust
first sentence of description: Effortlessly ethereal and romantic, this cutout-shoulder top is what dream closets are made of



(a) Metadata (b) Prompt 3 (c) Prompt 4

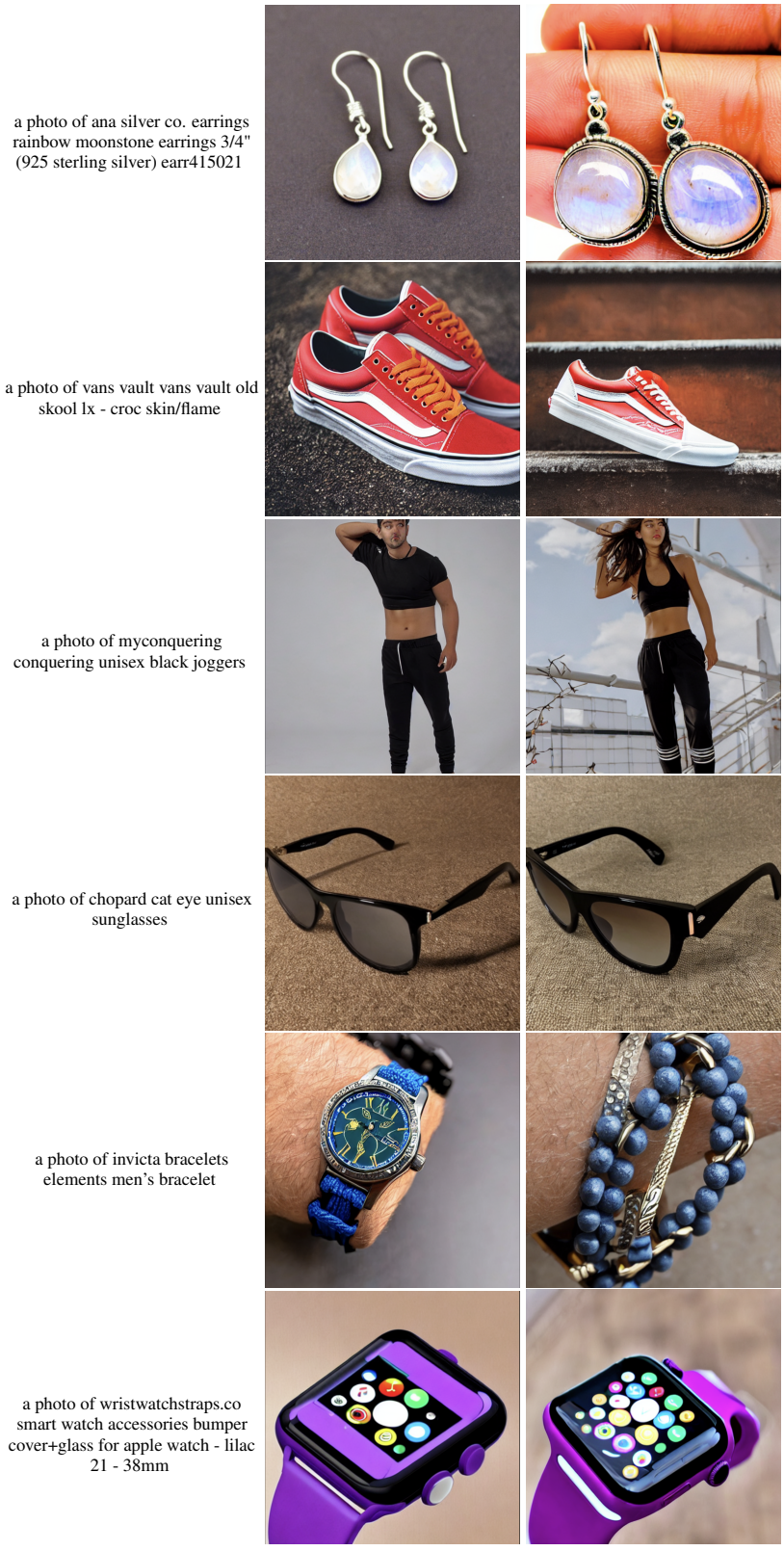
Figure Supp-7. Generated images with Vanilla SD model to determine prompt.

Prompt ID	Dataset	Prompt Structure
1	LGS	{brand} {title} in the style of e-commerce
	DeepFashion	{first sentence of description} in the style of e-commerce
2	LGS	{end_leaf} advertisement for a {title} from {brand}
	DeepFashion	{end_leaf} advertisement for a {first sentence of description}
3	LGS	{brand} {end_leaf} {title} {description}
	DeepFashion	{end_leaf} {description} {gender_category} {color}
4	LGS	a photo of {brand} {end_leaf} {title}, e-commerce
	DeepFashion	a photo of {end_leaf} {color} {first sentence of description}, e-commerce

Table Supp-3. Prompts evaluated for text-to-image generation experiment. Prompt structures varied slightly due to available metadata across datasets.

(a) Input Prompt (b) Vanilla SD (c) LGS-117

Figure Supp-8. Additional qualitative examples of the Vanilla SD vs LGS-117 fine-tuned SD model on DeepFashion InShop dataset.



(a) Input Prompt (b) Vanilla SD (c) LGS-117

Figure Supp-9. Additional qualitative examples of the Vanilla SD vs LGS-117 fine-tuned SD model on the LGS dataset.