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

Yatong Bai1∗ Utsav Garg2 Apaar Shanker2 Haoming Zhang2 Samyak Parajuli2
Erhan Bas2 Isidora Filipovic3 Amelia N. Chu3 Eugenia D. Fomitcheva3 Elliot Branson2
Aerin Kim2 Somayeh Sojoudi1 Kyunghyun Cho3

1University of California, Berkeley 2Scale AI 3New York University
∗Work done during internship at Scale. Correspondences to yatong_bai@berkeley.edu, aerinykim@gmail.com.

– Abstract –

Vision and vision-language applications of neural networks, 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 15 million image-caption pairs from publicly available e-commerce websites. When compared with existing general-domain datasets, the LGS images focus on the foreground object and have less complex backgrounds. Our experiments on LGS show that the classifiers trained on existing benchmark datasets do not readily generalize to e-commerce data, while specific self-supervised visual feature extractors can better generalize. 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 COCO [5]. This shortage of available public datasets can be attributed in part to the time and effort required to gather, clean, and annotate large 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-texts1 (Conceptual Captions [51]) for annotation sources. Nevertheless, social media data suffer from subjectivity. On the other hand, alt-texts can be unacceptably noisy, sometimes merely including uninformative texts such as “alt img”, as shown in Figure 1.

1Alt-texts are short descriptions of HTML website images. When an image cannot be rendered, the website displays its alt-text as a surrogate.
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. 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 uses 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 under the “BSD 3-Clause” license (also used in common datasets such as ImageNet), 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, unimodal 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.

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 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 resulting features is 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 to collect 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 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-

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Let’s Go Shopping (this paper)</td>
<td>14,847,764</td>
</tr>
<tr>
<td>YFCC100M (Yahoo)</td>
<td>100 million</td>
</tr>
<tr>
<td>RedCaps (University of Michigan)</td>
<td>12,011,111</td>
</tr>
<tr>
<td>Conceptual Captions 12M (Google)</td>
<td>12,423,374</td>
</tr>
<tr>
<td>WIT-English (Google)</td>
<td>5,500,746</td>
</tr>
<tr>
<td>Localized Narratives (Google)</td>
<td>849,000</td>
</tr>
<tr>
<td>COCO (Microsoft)</td>
<td>328,000</td>
</tr>
<tr>
<td>Visual Genome (Stanford)</td>
<td>108,077</td>
</tr>
<tr>
<td>CLIP (OpenAI)</td>
<td>400M</td>
</tr>
<tr>
<td>ALIGN (Google)</td>
<td>1.8B</td>
</tr>
</tbody>
</table>

Table 1. The instance count of LGS compared with existing bi-modal datasets.
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 mention the background. In contrast, e-commerce product thumbnails in LGS often depict only one inanimate 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 the words and phrases in LGS captions are diverse. For example, while LGS has around 3x more captions than COCO\textsuperscript{2}, its captions possess about 20x more uni-grams, bi-grams, and tri-grams, with more detailed

\begin{table}
\centering
\caption{Comparing the word count statistics of the LGS and COCO captions.}
\begin{tabular}{|c|c|c|c|c|c|}
\hline
Dataset & Min & Max & Mean & Median & Skew \\
\hline
LGS & 2 & 3642 & 89.58 & 67 & 3.44 \\
COCO & 5 & 50 & 10.56 & 10 & 2.76 \\
\hline
\end{tabular}
\end{table}

\begin{table}
\centering
\caption{The POS’s that occur at least ten times.}
\begin{tabular}{|c|c|c|c|c|}
\hline
Dataset & C. Nouns & P. Nouns & Adjectives & Verbs \\
\hline
LGS & 158,479 & 139,174 & 48,907 & 57,481 \\
COCO & 10,403 & 1,655 & 3,053 & 4,961 \\
\hline
\end{tabular}
\end{table}

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. Section 3.3 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.

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 informa-
tive 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 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 of 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 pre-processing 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 Classifiers 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
The classification accuracy of models trained on LGS shows that the LGS end leaves are well-separable.

<table>
<thead>
<tr>
<th>LGS Accuracy</th>
<th>LGS-117 from scratch</th>
<th>LGS-117 IN-pretrained (Top-1)</th>
<th>LGS-710 IN-pretrained (Top-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After linear probing</td>
<td>–</td>
<td>69.58 %</td>
<td>60.72 %</td>
</tr>
<tr>
<td>After fine-tuning</td>
<td>97.89 %</td>
<td>98.16 %</td>
<td>77.27 %</td>
</tr>
</tbody>
</table>

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”.

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Inception (↑)</th>
<th>FID (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet-1k</td>
<td>9.2930</td>
<td>114.60</td>
</tr>
<tr>
<td>IN pretrain→IN+LGS</td>
<td>9.1906</td>
<td>115.48</td>
</tr>
<tr>
<td>LGS</td>
<td>10.187</td>
<td>91.387</td>
</tr>
</tbody>
</table>

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 Section 3.4 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 Non-classification 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 performance 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 5. 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 Section 4.1.2 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).
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 7 shows that a two-phase ImageNet→LGS-710 weakly-supervised pre-training scheme produces features more suitable for fine-tuning on 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 in 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
Table 8. IC model performance of image-captioning task evaluated on different combinations of training and evaluation datasets.

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Test Set</th>
<th>METEOR (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGS-title</td>
<td>LGS-title</td>
<td>0.184</td>
</tr>
<tr>
<td>LGS-description</td>
<td>LGS-title</td>
<td>0.161</td>
</tr>
<tr>
<td>LGS-taxonomy</td>
<td>LGS-taxonomy</td>
<td>0.584</td>
</tr>
<tr>
<td>COCO</td>
<td>LGS-title</td>
<td>0.069</td>
</tr>
</tbody>
</table>

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

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;

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 8.

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) SD checkpoint.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set</th>
<th>FID (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>LGS Val</td>
<td>25.3498</td>
</tr>
<tr>
<td>Vanilla + LGS-117</td>
<td>LGS Val</td>
<td>24.1952</td>
</tr>
<tr>
<td>Vanilla</td>
<td>DeepFashion</td>
<td>62.9269</td>
</tr>
<tr>
<td>Vanilla + LGS-117</td>
<td>DeepFashion</td>
<td>74.0185</td>
</tr>
</tbody>
</table>
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. 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 ?? 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 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.

6. Acknowledgment

We would like to appreciate Matteo Bruno, James Caulkins, Xiaowen Dong, Peter Grindrod, Jack Hessel, Sean Holloway, Dorothy Nicholas, Janet Pierrehumbert, and Julian Winkler for their valuable help and fruitful discussions.

We also appreciate Baillie Gifford, the Institute for New Economic Thinking at the Oxford Martin School, and the UK Engineering and Physical Science Research Council for funding our work at the University of Oxford.

This work was supported in part through the NYU IT High-Performance Computing resources, services, and staff expertise.

References


[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, 6


[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. 14


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

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


[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. 15
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 Features learned on ImageNet and LGS 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 “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 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
Figure Supp-4. GradCam visualizations show that LGS classification models look for much more localized patterns.

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.

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 upsampled 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 ??). 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
Table Supp-2. The FID scores across prompts using a subset ($n = 5000$) of the LGS and DeepFashion InShop datasets.

<table>
<thead>
<tr>
<th>Prompt ID</th>
<th>Model</th>
<th>FID (↓)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LGS</td>
<td>DeepFashion</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Vanilla</td>
<td>40.4437</td>
<td>61.8519</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vanilla + LGS-117</td>
<td>42.7328</td>
<td>74.4327</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Vanilla</td>
<td>42.1081</td>
<td>63.2344</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vanilla + LGS-117</td>
<td>42.0529</td>
<td>77.7190</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Vanilla</td>
<td>36.7157</td>
<td>58.2189</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vanilla + LGS-117</td>
<td>36.1946</td>
<td>79.3607</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Vanilla</td>
<td>38.4101</td>
<td>62.9269</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vanilla + LGS-117</td>
<td>38.4100</td>
<td>74.0185</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Prompt ID</th>
<th>Dataset</th>
<th>Prompt Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LGS</td>
<td>{brand} {title} in the style of e-commerce</td>
</tr>
<tr>
<td></td>
<td>DeepFashion</td>
<td>{first sentence of description} in the style of e-commerce</td>
</tr>
<tr>
<td>2</td>
<td>LGS</td>
<td>{end_leaf} advertisement for a {title} from {brand}</td>
</tr>
<tr>
<td></td>
<td>DeepFashion</td>
<td>{end_leaf} advertisement for a {first sentence of description}</td>
</tr>
<tr>
<td>3</td>
<td>LGS</td>
<td>{brand} {end_leaf} {title} {description}</td>
</tr>
<tr>
<td></td>
<td>DeepFashion</td>
<td>{end_leaf} {description} {gender_category} {color}</td>
</tr>
<tr>
<td>4</td>
<td>LGS</td>
<td>a photo of {brand} {end_leaf} {title}, e-commerce</td>
</tr>
<tr>
<td></td>
<td>DeepFashion</td>
<td>a photo of {end_leaf} {color} {first sentence of description}, e-commerce</td>
</tr>
</tbody>
</table>

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.
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.

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.

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.

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).

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.

Figure Supp-7. Generated images with Vanilla SD model to determine prompt.
Figure Supp-8. Additional qualitative examples of the Vanilla SD vs LGS-117 fine-tuned SD model on DeepFashion InShop dataset.
Figure Supp-9. Additional qualitative examples of the Vanilla SD vs LGS-117 fine-tuned SD model on the LGS dataset.