Automated Product Description Generation for E-commerce via Vision-Language Model Fine-tuning

Wei Zhao
Department of Computer Science
Stanford University
wzhao18@stanford.edu

Xinyan He
Department of Computer Science
Stanford University
xinyanh@stanford.edu

Ericka Liu
Department of Computer Science
Stanford University
yilin23@stanford.edu

Abstract

The rapid growth of e-commerce has necessitated business owners to provide comprehensive and engaging descriptions for their products to facilitate smooth user shopping experience. However, manually crafting these product descriptions is labor-intensive and time-consuming. Recent advances in deep learning-based generative models offer a promising solution to automate this process. In this paper, we leverage state-of-the-art vision-language models, including CLIP, BLIP, BLIP-2, and OFA, to generate detailed and compelling product descriptions from images and metadata using the Amazon Berkeley Objects dataset. Our evaluation demonstrates significant improvements in the performance of the fine-tuned models over their pretrained versions across numerous metrics, with our best model achieving a two-order-of-magnitude increase in CIDEr score compared to the baseline approach.

1. Introduction

In recent years, the e-commerce sector has experienced substantial growth in scale. To facilitate smooth shopping experience, business owners are required to provide comprehensive information, including both visual displays and textual descriptions, for their products. Writing informative and appealing product descriptions is crucial as they help customers quickly understand product features and differentiate between various options. Traditionally, these product descriptions are crafted manually. Although this method ensures high-quality content, it is labor-intensive, time-consuming, and struggles to keep pace with the expanding variety of products.

Recent advances in deep learning-based generative models present a promising solution for generating new content from images and textual prompts. State-of-the-art vision-language models [7, 18] have demonstrated astonishing capabilities across a wide spectrum of multimodal tasks, including image captioning, which involves generating descriptions for images. However, applying existing image captioning models directly to product description generation faces two significant challenges. First, standard image captioning aims to generate sentences that factually describes the elements of a scene in an objective and neutral tone [15]. In contrast, effective product descriptions require an engaging and persuasive tone to captivate potential buyers. Second, while image captioning models focus solely on visual characteristics, it is essential to include product metadata in product descriptions, such as manufacturing details and material composition. This requirement further underscores the inadequacy of existing image captioning models for generating effective product descriptions.

In this work, we explore the adaptation of deep learning models for the task of generating product descriptions for e-commerce platforms. Our goal is to build models that take product images and metadata of products as input and output informative and engaging descriptions for them. To accomplish this task, we select four widely used vision-language models as our base models, CLIP [12], BLIP [8], BLIP-2 [7], and OFA [18]. These models are fine-tuned using the Amazon Berkeley Objects (ABO) dataset [2], which includes information on more than 100k products listed on the Amazon platform. We provide a detailed overview of our fine-tuning implementation for each base model in section 4. In addition, we highlight our training strategies to mitigate the substantial resource requirements for large-scale model fine-tuning in section 5, including the comparison between Parameter Freezing and Low-Rank Adaptation (LoRA) [6]. In section 6, we evaluate the effectiveness of our fine-tuned models using a range of metrics, including BLEU [11], CIDEr [17], METEOR [1], and ROUGE-L [9]. Notably, our best-performing model, fine-tuned from the OFA base model, achieves a remarkable two-order-of-magnitude increase in CIDEr score compared to the baseline approach. This significant improvement underscores the potential of our approach in enhancing the performance of pretrained vision-language models for the task of product description generation.

2. Related Work

Vision Pretraining & Language Pretraining. In natural language processing (NLP), learning universal language representations through pretraining has significantly enhanced performance across numerous tasks, including natural language inference, named entity recognition, and question answering [3]. Pretrained language models, such as BERT [3] and GPT [14], leverage vast corpora of text to train large Transformer-based models [16] to learn sophisticated linguistic structures and nuanced sentence semantics in an unsupervised manner. Pretrained models can subsequently be fine-tuned using relatively small, task-specific datasets to render robust performance in a variety of downstream tasks. This pretrain-finetune paradigm not only improves the efficiency of training models for specific tasks by utilizing shared pretrained models, but also enhances performance with the transfer of knowledge learned from large-scale dataset.

Inspired by NLP's success through langugae pretraining, numerous efforts have sought to adapt the unsupervised pretraining techniques of transformers to the domain of computer vision. Vision Transformer (ViT) [4] adapts the Transformer architecture for visual data by segmenting an image into patches and then providing the sequence of linear embeddings of these patches as inputs to a Transformer model, treating each image patch as a word token in language models. Pretrained on large image datasets, these models excel in tasks like image recognition and classification, often surpassing the performance of traditional convolutional neural networks (CNNs) like ResNet [5].

Vision-language Pretraining. Vision-language pretraining aims to learn joint representations for both vision and language, enabling models to perform multi-modal tasks such as image captioning, visual question answering, and text-to-image generation. CLIP [13] employs contrastive learning to predict correct image-caption pairs, leveraging large-scale internet data. Florence [20] adapts contrastively pre-trained models to more downstream tasks with task-specific adaptations. BLIP [12] unifies the handling of vision-language understanding and text-generation tasks with the Multimodal mixture of Encoder-Decoder architecture (MED). OFA [18] extends the generation capability to include text-to-image generation with a task and modality-agnostic sequence-to-sequence framework. Furthermore, BLIP-2 [7] was proposed to address the soaring cost of vision-language pretraining by leveraging off-the-shelf frozen pre-trained image encoders and frozen large language models. It bridges the gap between the two modalities gap with a lightweight module known as Querying Transformer (Q-Former). This method demonstrated competitive performance in vision-language tasks while requiring significantly fewer trainable parameters.

3. Dataset

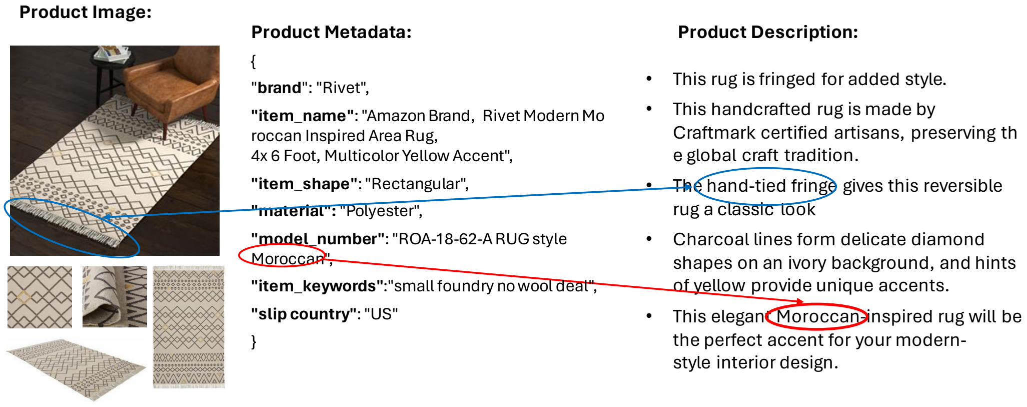
The dataset utilized in this study is the Amazon Berkeley Objects (ABO) Dataset [2], a comprehensive collection encompassing 147,702 Amazon product entries. Each entry includes a set of catalogue images and associated metadata as illustrated in Figure 1. A typical entry within the ABO Dataset comprises a sequence of images that showcase the product from various perspectives, providing a detailed visual representation of its appearance and features. Alongside these images, the dataset includes product metadata presented in a structured JSON format, which encapsulates essential attributes such as dimensions, materials, and manufacturer information. This metadata offers valuable textual information that complements the visual content and aids in the comprehensive understanding of each product. The corresponding product descriptions are provided on the right side of the figure, formatted as a series of bullet points. These descriptions emphasize the key features and specifications of the product. Notably, the generation of these descriptions relies on synthesizing information from both the visual attributes observable in the images and the textual details contained within the metadata. This observation underscores the importance of leveraging both visual and textual data modalities to accurately reconstruct the original product descriptions.

4. Methods

We investigate the adaptation of pretrained vision-language models for the task of generating product descriptions from images and structured metadata. Specifically, we explore four widely used models: CLIP [12], BLIP [8], BLIP-2 [7], and OFA [18]. In this section, we begin with a high-level overview of the design and architectures of these base models, along with the mechanism of the LoRA method. Then, we delve into the implementation details. This includes the preprocessing of input data and the fine-tuning procedures applied to each individual model.

4.1. Base Models and Techniques

CLIP. CLIP leverages contrastive learning to jointly pretrain both a vision encoder and a text encoder. The vision encoder can be constructed using either CNNs such as ResNet or Transformer-based architectures like Vision Transformer (ViT). The text encoder is built upon a GPT-style Transformer architecture. During the training phase, CLIP trains the image and text encoders to predict



Product Metadata:

```
{
"brand": "Rivet",
"item_name": "Amazon Brand,  Rivet Modern Mo
roccan Inspired Area Rug,
4x 6 Foot, Multicolor Yellow Accent",
"item_shape": "Rectangular",
"material": "Polyester",
"model_number": "ROA-18-62-A RUG style
Moroccan",
"item_keywords":"small foundry no wool deal",
"slip country": "US"
}
```

Product Description:

- This rug is fringed for added style.
- This handcrafted rug is made by Craftmark certified artisans, preserving th e global craft tradition.
- The hand-tied fringe gives this reversible rug a classic look
- Charcoal lines form delicate diamond shapes on an ivory background, and hints of yellow provide unique accents.
- This elegant Moroccan-inspired rug will be the perfect accent for your modern-style interior design.

Figure 1: Example of product image, metadata, and descriptions from the ABO dataset. Descriptions such as 'hand-tied fringe' are inferred from the image, while others like 'Moroccan style' are inferred from the metadata.

the correct pairings within a batch of (image, text) training examples. In the testing phase, the model synthesizes a zero-shot linear classifier by embedding the class names or descriptions of the target dataset, enabling it to generalize to new tasks without additional training.

BLIP. BLIP employs a multimodal mixture of encoder-decoder (MED) pre-training scheme, which comprises three key components: an image-grounded text encoder, an image encoder, and a multimodal decoder. The image-grounded text encoder is responsible for learning textual descriptions in the context of their corresponding images, effectively capturing the semantic relationships between the two modalities. Meanwhile, the image encoder focuses on learning rich visual representations that encapsulate the salient features and attributes of the input images. Lastly, the multimodal decoder generates textual outputs based on the combined visual and textual inputs provided by the image and text encoders.

BLIP-2. Unlike prior works which jointly trains vision and language components, BLIP-2 leverages frozen pre-trained vision encoders and language models to reduce training cost. This is achieved with the introduction of the query-based transformer (Q-Former), which acts as a bridge between the visual and textual modalities. Q-Former transforms the visual features into query embeddings through a series of attention mechanisms. These query embeddings then serve as the input to the language model, enabling it to generate contextual text outputs, such as captions for the images.

OFA. OFA is a transformer-based sequence-to-sequence model designed to tackle a wide range of crossmodal and unimodal tasks within a single framework. OFA represents data of various modalities, such as texts, images, and objects (in images), as tokens in a unified vocabulary. This is accomplished by utilizing CNNs to transform images into patch features and byte-pair encoding (BPE) to encode text sequences as subword sequences. The architecture of OFA follows the encoder-decoder paradigm, where both the encoder and the decoder are composed of stacks of Transformer layers. Each encoder layer consists of a self-attention module and a feed-forward network (FFN), while each decoder layer includes a self-attention block, an FFN, and a cross-attention mechanism to establish the connection between the decoder and the encoder output representations.

LoRA. LoRA is a fine-tuning technique which significantly reduces the computational and memory requirement associated with fine-tuning by injecting trainable rank decomposition matrices into specific layers of a pre-trained model. Specifically, for each selected layer of the model, LoRA injects low-rank matrices that decompose the original weight matrix. During the fine-tuning phase, the original weight matrices of the model remain unchanged. Instead, the outputs from the low-rank matrices are used to adjust the activations of the layers in the model. The dimensions of the low-rank matrices can be tuned to balance between expressiveness and computational efficiency.

4.2. Implementation Details

Data Preprocessing. Given the objective of the study to generate product descriptions from image and textual metadata, we first apply a filtering step to eliminate data points that either do not contain an image or do not have at least one product description written in English. After this refinement, the resulting dataset comprises 113,203 samples, with an average of 4.8 images per product. We

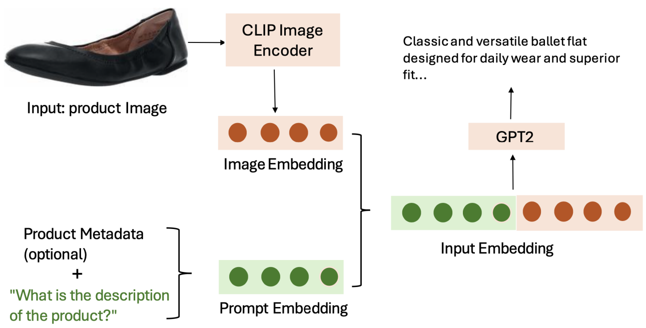

CLIP Image Encoder
Input: product Image
Image Embedding
Classic and versatile ballet flat designed for daily wear and superior fit...
GPT2
Input Embedding
Product Metadata (optional)
+
"What is the description of the product?"
Prompt Embedding

Figure 2: Overview of CLIP-GPT2 architecture

parse the originally JSON-formatted metadata into a textual format to allow the model for easier generalization and adaptability to data that may not be presented in JSON format. Furthermore, we remove certain keywords from the metadata, such as $language\_tag$, which is used to indicate the language of the corresponding value. This removal streamlines the metadata and focuses on the essential information required for generating product descriptions. Lastly, to ensure compatibility with the English vocabulary used in the tokenization process, we apply an additional filtering step to remove all information that is not written in English.

CLIP-GPT2. Inspired by ClipCap [10], we introduce a new streamlined model called CLIP-GPT2, which integrates CLIP's image encoder with the GPT2 language model. An overview of the CLIP-GPT2 model in presented figure 2. As shown, metadata is prepended to GPT-2's prompt as a prefix. In addition, the model maps the visual embeddings of CLIP's image encoder to the language model's latent space, creating a learnable visual prefix. This prefix serves as a guide for the language model to generate relevant descriptions of the image content. For implementation, we use pretrained weights for GPT2 from ”openai-community/gpt2” and image encoder from ”clip-vit-base-patch32”. In total, CLIP-GPT2 comprises approximately 275 million parameters. Unlike the original ClipCap framework, which freezes the image encoder to facilitate efficient parameter tuning, we choose to fine-tune both the image encoder and the language model. This decision is driven by the difference between the ABO dataset, which consists of commercial product images, and the datasets on which these models were originally pretrained.

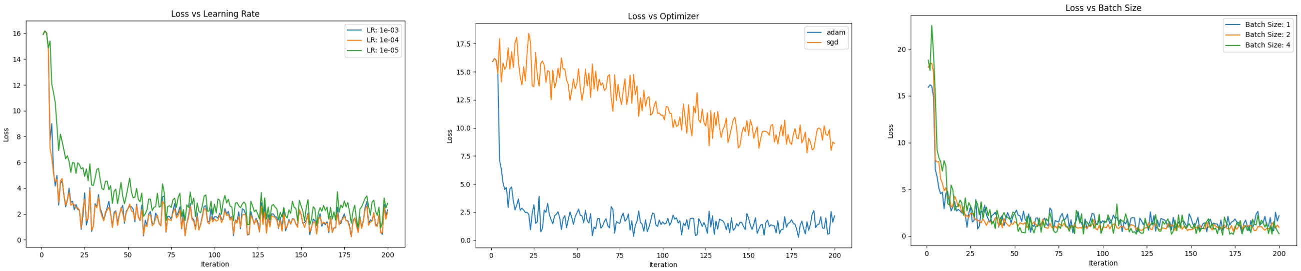
BLIP Fine-tuning. Since BLIP employs different models tailored to specific tasks, for our use case, we utilize the BLIPForQuestionAnswering model, which is used for the task of visual question answering (VQA). The model takes an image and a prompt as input and generates an answer accordingly. We initialize the model with pretrained weights from the ”Salesforce/blip-vqa-base” checkpoint. To prepare the input data for the model, we tokenize the parsed metadata and the target output separately. These tokenized inputs are then passed to the model, which processes the image and the prompt using its integrated visual and textual encoders.

BLIP-2 Fine-tuning. For BLIP-2, we employ the smallest variation of the model, namely ”Salesforce/blip2-opt-2.7b”, which comprises a Vision Transformer (ViT) for visual encoding and an OPT-2.7b language model for text generation. The challenge associated with this model is its substantial model size of 3.1 billion parameters, occupying 14.43 GB of memory for the weights alone. A naive approach to training the model using the Adam optimizer would demand 57.72 GB of memory. To reduce memory usage, we explored parameter freezing and Low-Rank Adaptation (LoRA). Parameter freezing, as used in the original BLIP-2 paper, involves freezing the vision encoder and language model weights during fine-tuning. Alternatively, LoRA reduces trainable parameters by introducing low-rank decomposition matrices. We offer a detailed comparison of these two approaches in section 5. Given the decoder-only architecture of the OPT language model, we concatenate the input metadata with the corresponding output descriptions. The tokenized representation of this concatenated string is used as both the input and the labels.

OFA Fine-tuning. For fine-tuning the OFA model, we initialize the model and the tokenizer with pretrained weights from the ”OFA-Sys/ofa-large” checkpoint. Given the encoder-decoder architecture of OFA, we utilize the image and metadata as inputs to the encoder, while the ground truth descriptions serve as targets for the decoder. For image processing, we preprocess the images by resizing and cropping them to a resolution of 256 × 256 pixels, which are then transformed into 16 × 16 patches by the OFA tokenizer. One notable aspect of the OFA model is its task-agnostic design, meaning that there is no task-specific projection layer at the end of the model. We use the instruction fine-tuning technique by appending the question ’What is the description of the product?’ to the prompt, guiding the model to generate product descriptions. To accelerate the training process, we freeze the encoder module and adopt fp16 training functionality provided by the Trainer interface from the Huggingface Transformers library [19].

5. Training Details

We split the dataset into train, validation, and test sets with a ratio of 7:2:1. Before starting the full training process, we use a small subset of the training dataset



(a) Loss with different learning rates

(b) Loss with different optimizers

(c) Loss with different batch sizes

Figure 3: Hyperparameter Tuning for OFA

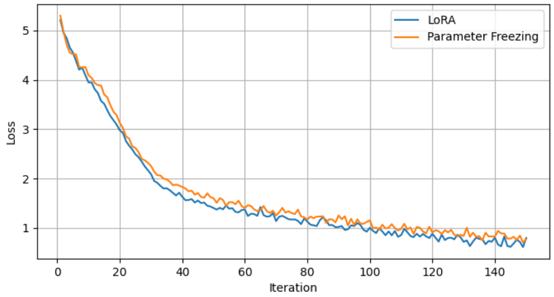


Figure 4: BLIP-2 loss curve for 10 epochs using LoRA and Parameter Freezing

with 1,000 examples to verify the correctness of our implementation and estimate the training cost by observing the loss curve. Once the models exhibit clear signs of learning, we proceed to a small-scale hyperparameter tuning phase. The goal of this phase is to identify the optimal hyperparameters for the fastest convergence rate for each model. We illustrate our approach by presenting the hyperparameter tuning process for the OFA model as a representative example. Additionally, we highlight a comparative experiment evaluating the effectiveness of LoRA and parameter freezing in the context of fine-tuning BLIP-2.

Hyperparameter Tuning. Figure 3 presents the results of our comprehensive hyperparameter tuning process for the OFA model. In 3a, we evaluate the impact of different learning rates on the model's performance. We explore three learning rate values: 1e-03, 1e-04, and 1e-05. Our findings reveal that while a learning rate of 1e-03 achieves rapid convergence, it results in high fluctuations in the loss curve. In contrast, a learning rate of 1e-04 exhibits a more stable convergence pattern with fewer fluctuations, while consistently maintaining a lower overall loss compared to the 1e-05 setting. Based on these observations, we select 1e-04 as our final learning rate, as it strikes a balance between convergence rate and stability. Furthermore, figure 3b presents a comparative analysis of the loss curves obtained using the Adam and SGD optimizers. The results clearly demonstrate the superiority of the Adam optimizer, as it achieves a significantly lower and more stable loss throughout the 200 iterations compared to SGD. Consequently, we opt for the Adam optimizer in our training configuration. Furthermore, we investigate the effects of different batch sizes on training stability and convergence. We evaluate batch sizes of 1, 2, and 4, as shown in Figure 3c. All three batch sizes exhibit a similar trend, characterized by a rapid decrease in loss during the initial iterations, followed by a more gradual decline as training progresses. Based on these findings, we choose a batch size of 4 as it maximizes training throughput.

LoRA vs. Parameter Freezing for BLIP-2. Given the substantial computational resources required to train the BLIP-2 model, we begin with a small-scale experiment using 1000 randomly selected samples from the training dataset to estimate the rate of convergence and training time. For the LoRA approach, we set the rank to 256, making 2.19% of the total parameters trainable. For the Parameter Freezing technique, we freeze all layers in both the vision encoder and the language model, leaving 2.86% of the total parameters trainable. This approach ensures a comparable scale of trainable parameters for both techniques, allowing us to evaluate their relative efficiency and effectiveness under similar conditions. Figure 4 illustrates the training curves over 10 epochs using the LoRA and Parameter Freezing approaches. Both techniques exhibit similar rates of loss reduction, with LoRA demonstrating a slightly faster convergence. However, at the 10th epoch, the loss continues to decrease noticeably, indicating that the model is still far from reaching convergence. Using the full training dataset, the time per epoch for LoRA and Parameter Freezing is approximately 10 and 9 hours, respectively. Based on these observations, we estimate that the training cost for both approaches would exceed the available resources allocated for this study. Consequently, we decide not to proceed with the BLIP-2 model in further investigations.

	# Trainable Params↑	CIDEr↑	BLEU-1↑	BLEU-4↑	ROUGE-L↑	METEOR↑
Metadata-Only		0.85	5.84	1.48	7.63	10.7
CLIP-GPT2-Baseline	276M	2.02e-2	5.79	0.15	4.76	10.50
CLIP-GPT2-Fine-tuned	276M	1.01	13.89	5.78	30.89	20.45
Blip-Baseline	384M	2.10	2.06e-3	6.34e-4	6.42	3.40
Blip-Fine-tuned	384M	7.73	48.11	36.32	57.56	59.90
OFA-Baseline	472M	7.93e-03	1.98e-03	9.14e-4	1.85	1.31
OFA-Fine-tuned	472M	**473.38**	**74.37**	**68.71**	**74.61**	**74.46**

Table 1: Performance Evaluation on ABO dataset

6. Evaluation

6.1. Methodology

Baselines. We establish two fundamental baselines to evaluate the effectiveness of the fine-tuned models for the task of product description generation. The first baseline involves directly utilizing the parsed metadata as the product description. Although this approach may lack the cohesiveness of natural language, it is selected as a baseline due to its comprehensive coverage of product information. The second baseline leverages the aforementioned pretrained vision-language models CLIP-GPT2, BLIP, and OFA) to generate textual descriptions based on both the visual content of the product image and the parsed metadata.

Metrics. We compare the performance of the fine-tuned models against the baseline approaches on the ABO datasets using several evaluation metrics, including BLEU [11], CIDEr [17], METEOR [1], ROUGE-L [9]. Given the similarity between the task of product description generation and image captioning, we prioritize CIDEr as the primary evaluation metric. CIDEr measures how well the generated descriptions match human-annotated references, emphasizing the importance of relevance and specificity. BLEU, METEOR, and ROUGE-L are also used to provide a comprehensive evaluation of the generated descriptions' fluency, precision, and recall.

6.2. End-to-end Results

Quantitative Evaluation. We conducted a comprehensive comparative analysis of various baseline and fine-tuned models, as summarized in Table 1. The results demonstrate that fine-tuning significantly improves performance of the models compared to their baseline versions, which exhibited suboptimal performance compared to the metadata-only baseline. Notably, the OFA model, despite requiring a larger number of trainable parameters (472M) compared to the other models (384M and 276M for BLIP and CLIP-GPT2 respectively), delivers superior results across all five evaluation metrics. In particular, the OFA model achieves a CIDEr score of 473.38 in the test dataset, which is two orders of magnitude higher than the best baseline approach (2.10). This substantial improvement in the CIDEr score highlights the OFA model's ability to generate product descriptions that closely mimic the targeted outputs. While the BLIP model maintains comparable scores to the OFA model in terms of BLEU-1, BLEU-4, ROUGE-L, and METEOR, its CIDEr score is notably lower at 7.73. This disparity arises because CIDEr emphasizes the use of unique and less common n-grams, which are more informative about the specific characteristics of the product. The BLIP model tends to generate more generic descriptions that lack the distinctiveness required to achieve a high CIDEr score, whereas the OFA model excels at producing detailed and unique descriptions that capture the essence of each product.

Qualitative Evaluation. To elucidate the efficacy of various models in crafting appealing product descriptions, we present a qualitative demonstration in Table 2. This table showcases detailed product input data and the corresponding output descriptions generated by the OFA, BLIP, and CLIP-GPT2 models. According to the table, CLIP-GPT2 produces the poorest quality captions. The model identifies keywords but fails to construct coherent and meaningful sentences, often resorting to placeholder text or generic phrases. This suboptimal performance could be attributed to the integration of CLIP and GPT-2, which requires mapping visual features from CLIP to a format compatible with GPT-2. Unlike OFA and BLIP, which are specifically designed and trained for vision-language tasks, this mapping may result in a loss of critical information and less effective integration between the visual and linguistic components, leading to incoherent sentence outputs. In contrast, OFA generates high-quality descriptions similar to ground truth labels. Notably, OFA demonstrates the ability to capture numerical information directly from images. For instance, it accurately identifies the volume of the syrup as 15.25 ounces in the first image, even though this informa-

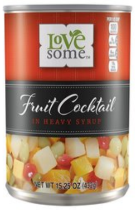
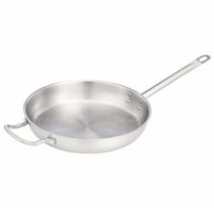
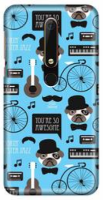
Images			
Metadata	{ ”brand”: ”LoVesome”, ”product_type” : ”grocery”, ”country”: ”US”, ”marketplace”: ”AmazonFresh”, ”domain_name”: ”fresh.amazon.com” }	{”height”: 3.54, ”length”: 25, ”width”: 12.32, ”unit”: ”inches”, ”color”: ”silver”, ”item_name”: ”Amazon Commercial 30 cm Stainless Steel Aluminium-Clad Frying Pan with Helper Handle”}	{”brand”: ”Amazon Brand - Solimo”, ”color”: ”multi-colored”, ”item_name”: ”Amazon Brand - Solimo Designer Bicycle Pattern Design 3D Printed Hard Back Case Mobile Cover for Nokia 6 (2018)”}
Ground Truth	- Brought to you by Whole Foods Market	- Measures 63 x 32 x 9 cm - 30 cm frying pan with a wide, flat shape and stainless-steel cooking surface, perfect for searing, sautéing, browning and more - designed for residential and commercial use	- 31.5”W x 16.54”D x 31.5”H - Pine, MDF and glass, Metal - The white finish and metal pulls add a modern element to the design. Geometric door trim gives it a graceful style that will go with most any decor.
OFA	- Brought to you by Whole Foods Market - One 15.25-Ounce can of Looseome Fruit Cocktail Heavy Syrup - Satisfaction Guarantee	- 7-inch stainless steel saute pan for cooking and baking - Made of heavy-gauge 18/8 stainless steel - reinforced edges for rugged impact-resistant strength and to maintain shape of pan	- 31.5”W x 16.54”D x 31.5”H - Pine, MDF and glass, Metal - The white finish and metal pulls add a modern element to the design. Geometric door trim gives it a graceful style that will go with most any decor.
BLIP	- brought to you by whole foods market. - our organic low sodium chicken broth is made from slow, simmered organic chicken, as well as the classic flavor combination of organic carrots and onions.	- sterling silver necklace featuring polished round disk initial k pendant - 18’ cable chain with spring - ring clasp - great for everyday wear - these silver pieces are built for longevity.	- snug fit for mobile, with perfect cut - outs for volume buttons, audio and charging ports - easy to put & take off with perfect cutouts for volume buttons, audio & charging ports. - stylish design and appearance, express your unique personality.
CLIP-GPT2	- for a products in Whole Foods Certified - no.1 ingredients.; - grab every meal the day for what is exceptional - products of in compromises has at least are chickenKitchen joy	- with: (.) from the use stainless steel in easy on corrosion coating; - Bakelite Dish washerable	- soft rubber snug interior feel smooth - surfaces lightweight without scratching - comfortable” inches scratches surfaceable phone Case fit in corners

Table 2: Selected product information and generated captions for different models.

tion is not present in the textual metadata. For the phone case example, OFA produces a description identical to the target description. This can be attributed to the large number of phone case products in the training dataset, which have similar descriptions that only vary in specific details such as dimensions. The consistency in the quality of generated descriptions across different product types highlights the robustness and reliability of the fine-tuned OFA model. The results produced by BLIP are also commendable. However, for the second image, it incorrectly identifies the object as a necklace when it is, in fact, a frying pan. This suggests that while BLIP can generate high-quality descriptions, it struggles with distinguishing between visually similar objects.

Despite OFA's effective generation of descriptions that closely resemble the target labels, there are known limitations in the current approach. Table 3 illustrates some failure examples of the OFA model, where it generates mean-

Images		
Generated	- AmazonBasics No expanded, Black - An Amazon Brand - Content coming soon	- This is a placeholder

Table 3: Failure Examples of OFA

ingless descriptions such as ”content coming soon” and ”This is a placeholder” for the two selected products. These failures are primarily due to the presence of noisy samples in the training dataset, which use placeholder sentences as descriptions. Ideally, the model should be able to identify and disregard these noisy examples, allowing it to generate meaningful descriptions despite their presence.

7. Conclusion

In this work, we explore fine-tuning state-of-the-art vision-language models to automate the generation of product descriptions for e-commerce platforms. Our results demonstrate significant improvements in the fine-tuned models over their pretrained versions. Notably, the fine-tuned OFA achieves the highest scores across all evaluation metrics. Despite the model’s success in producing similar outputs as the target descriptions, we have also identified its deficiency when certain queries are corrupted by noisy examples in the training dataset. Future work could focus on mitigating the impact of noisy data, scaling the approach to a broader range of products and categories, and integrating additional contextual information to enhance the quality and relevance of descriptions. These improvements can further improve the quality of the generated product descriptions and reduce the labor of manually crafting descriptions by business owners of e-commerce platforms.

8. Acknowledgement

Special thanks to our mentor: Saumya Goyal. We do not have external collaborators and we are not sharing projects. Contribution: We equally contributed to the writing. Ericka: Focused on implementing OFA baseline and fine-tuning, as well as hyperparameter tuning. Xinyan: Worked on implementing CLIP and BLIP-2 baseline and fine-tuning, as well as quantitative evaluation. Wei: Contributed to the implementation and fine-tuning of BLIP and OFA, and the LoRA integration for BLIP-2, as well as the qualitative evaluation.

References

[1] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[2] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Yago Vicente, T. Dideriksen, H. Arora, M. Guillaumin, and J. Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.

[7] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[8] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. volume abs/2201.12086, 2022.

[9] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[10] R. Mokady, A. Hertz, and A. H. Bermano. Clipcap: CLIP prefix for image captioning. volume abs/2111.09734, 2021.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL ’02, page 311–318, USA, 2002. Association for Computational Linguistics.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. volume abs/2103.00020, 2021.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *OpenAI blog*, 1(8):1–12, 2018.

[15] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. *CVPR*, 2019.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[17] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation, 2015.

[18] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022.

[19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[20] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang. Florence: A new foundation model for computer vision, 2021.