VIT-Pro: Visual Instruction Tuning for Product Images

Vishnu Prabhakaran[1], Purav Aggarwal[1], Vishruit Kulshreshtha[1], Arunita Das[1], Sahini Venkata Sitaram Sruti[1,2], Anoop Saladi[1]
[1]Amazon, India, [2]Indian Institute of Technology, Patna
{visprab,aggap,kulshrev,arunita,ssruti,saladias}@amazon.com

Abstract

General vision-language models (VLMs) trained on web data struggle to understand and converse about real-world e-commerce product images. We propose a cost-efficient approach for collecting training data to train a generative VLM for e-commerce product images. The key idea is to leverage large-scale, loosely-coupled image-text pairs from e-commerce stores, use a pre-trained LLM to generate multi-modal instruction-following data, and fine-tune a general vision-language model using LoRA. Our instruction-finetuned model, VIT-Pro, can understand and respond to queries about product images, covering diverse concepts and tasks. VIT-Pro outperforms several general-purpose VLMs on multiple vision tasks in the e-commerce domain.

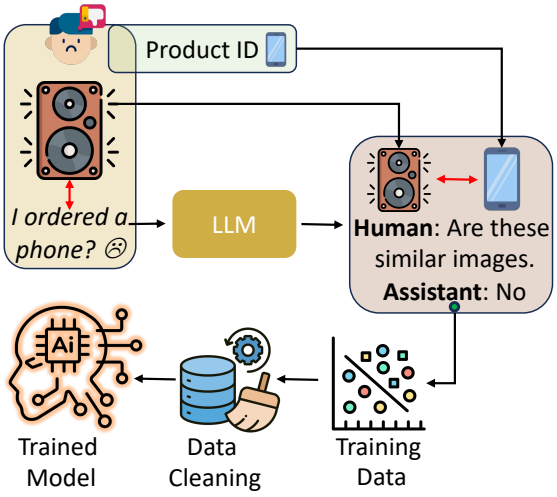

Figure 1: Illustration of how multi-modal feedbacks are collected, processed and refined to curate training data for training a model.

1 Introduction

The e-commerce domain inherently operates at the intersection of visual and textual data. From high-resolution product images and packaging photos to detailed customer feedbacks provided during return/refund claims, the interplay between these modalities is central to ensuring smooth operations and customer satisfaction. This multi-modal nature of data is pivotal in scenarios like verifying product authenticity, monitoring quality control, and resolving customer grievances effectively. However, the sheer volume of such data, generated across stages of the logistics chain—packaging, shipping, delivery, and post-delivery—poses a significant challenge. Efficiently leveraging this wealth of multi-modal information is critical for scaling operations while maintaining accuracy and customer trust. Currently, manual investigations to address multi-modal customer queries, such as verifying product quality and delivery issues, are the standard practice but lack scalability. Automating these investigations requires robust multi-modal systems that can precisely analyze visual and textual data together.

While general-purpose Vision-Language Models (VLMs) are proficient in handling diverse tasks, they often lack the nuanced understanding required for domain-specific applications in the e-commerce sector. These applications include accurately recognizing and differentiating between similar products among a vast collection, extracting specific product attributes from images and descriptions, understanding product compatibility and accessorizing requirements, and assessing product quality and detecting damages or defects based on images. An additional challenge arises from real-world ("in-the-wild") images, as most images (apart from catalog images) are non-standard, with varying viewpoints, partially visible regions, occluded parts, and poor quality. To address these challenges, e-commerce stores may need to develop specialized VLMs tailored to their specific domains. However, the development of such systems is hindered by

the unavailability of domain-specific multi-modal datasets. Addressing this data bottleneck is crucial to enabling automation at scale.

To bridge this gap, we propose a scalable framework for curating multi-modal instruction-following datasets tailored to the e-commerce domain (illustrated in Figure 1). This approach leverages readily available customer feedbacks, product catalog and associated images to transform them into rich instruction-following dataset using a pre-trained LLM. To ensure quality, we employ robust cleaning techniques, including attention-guided data validation, to filter irrelevant or noisy samples. The curated dataset facilitates the fine-tuning of vision-language models, equipping them with e-commerce-specific capabilities. Our work makes the following key contributions:

- *E-Commerce Multi-Modal Instruction-Following Data:* We introduce a novel data generation strategy that transforms weakly associated image-text pairs from existing sources into a high-quality, multi-modal instruction-following dataset. This dataset, comprising 1.4M unique samples across diverse e-commerce tasks, is generated without manual annotation efforts.
- *Visual Attention Guided Data Refinement:* We propose a novel and effective method that uses transformer attention maps to compute visual grounding scores, allowing us to filter out samples with poorly grounded text segments.
- *VIT-Pro:* We present VIT-Pro, a multi-task multi-modal model fine-tuned using the curated dataset which is adapted to the e-commerce domain and demonstrate superior performance as compared to other open-source and commercially available visual language models for e-commerce tasks.

2 Related Work

Vision-Language Modelling for E-commerce has been studied and experimented in several existing works (Fu et al., 2022; Khandelwal et al., 2023; Jia et al., 2023). However, most of these works are targeted towards visual question answering tasks for attribute extraction, catalog quality improvement, etc. using high-quality product catalog images and texts. Consequently, these datasets and models are not scalable to other challenging tasks in the e-commerce domain involving in-the-wild product images (as discussed in section 1). Compared to these existing works, ours is a pioneering attempt towards building a e-commerce domain specific VLM that can answer open questions in the wild on real-world images and tasks applicable at various stages in the product order life cycle. More recently, Visual Instruction Tuning has proven to be a promising approach to enable models to follow diverse user instructions involving visual content. Several open-sourced instruction-tuned models, including InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023), IDEFICS2 (Laurençon et al., 2024), Qwen-VL (Bai et al., 2023) and larger propriety models, such as ClaudeV3 (Anthropic, 2024), GPT4V (OpenAI, 2023), achieve competitive performance on real-world tasks (VisIT-Bench (Bitton et al., 2023)). However, their zero-shot performance on e-commerce domain remains limited, because the e-commerce image-text pairs and tasks are drastically different from public datasets. Our work, inspired by these existing works, extends multi-modal instruction-tuning to the e-commerce domain by efficiently constructing an LLM-assisted instruction-following data from disparate sources for building a multi-task VLM with extended domain-specific capabilities.

3 Vision-Language Instruction Tuning

Foundational VLMs for e-commerce require domain-specific capabilities such as fine-grained product recognition, product counting, localization, attribute extraction, damage/defect detection, comparison of visually similar products, visual reasoning, multi-modal decisioning, etc. To impart this knowledge and instruction-following ability, we require training on a rich multi-task dataset meticulously crafted with distinct instructions and responses. To achieve this, we present techniques from both data and model perspective in this section.

3.1 E-commerce Visual Instruction Data

With an aim to collect a diverse and representative set of e-commerce concepts and products in the form of image-text pairs, we sample 4M image-text pairs from most common data sources available in e-commerce stores: product catalog, customer complaints and customer return/refund claim conversations, etc. We removed image-text pairs (≈20%) with a low CLIP score. However, the key challenge is that the raw data corpus is not directly

Context Type 1: Chat Transcripts / Feedback

Agent: Hello my name is [NAME] how may I help you?
Customer: [Greets] sir. I have a defective damaged item sir
Customer: This [BRAND] package torn at delivery
Agent: [NAME] Check [UCI] please [NAME] please attach images [NAME]
(....remaining transcript omitted for brevity)

Context Type 2: Product Information

[BRAND] with Double Power Jasmine and Rose Detergent Washing Powder - 1 kg

---

Generated Question-Answer Pairs:

Task: Product Recognition
Q: What's the product present in the picture ?
A: The image shows a [BRAND] Double Power detergent powder.

Task: Damage/Defect Detection
Q: Is the product damaged ?
A: Yes, the packet seems to be torn open and the detergent powder has spilled out.

Task: Product Matching
Q: Does the product in the image match to the Product Title: "[BRAND] Matic Liquid Detergent 3.2L Top Load Washing Machine"?
A: The image shows a package of [BRAND] powder detergent with "Double Power" as an additional feature, while the product tile is of a liquid detergent ... <omitted>

Task: Attribute Extraction
Q: Please use the image to extract product attributes in a JSON format.
A: {"ProductType": "Detergent Washing Powder", "Name": "[BRAND] Plus with Double Power", "Brand": "[BRAND]", "Color": "Unanswerable", "Count": "1", "Size": "Unanswerable", "ExpiryDate": "Unanswerable"}

Table 1: An example illustrating the instruction-following data. The top block shows the redacted contexts (customer chat transcripts / feedbacks / product information) used to prompt the LLM, and the bottom block shows the generated question-answer pairs. The image is displayed only for reference here and is not used to prompt the LLM.

usable, as the raw text in the form of chat/phone transcripts, etc. can be unstructured and irrelevant (more sentimental content).

Motivated by recent success of LLMs in data preparation and processing (Touvron et al., 2023; Chen et al., 2023; Li et al., 2023), we leverage open source LLMs (internally hosted APIs) for generating multi-modal instruction-following data, from existing raw data corpus. Specifically, given raw texts from customer feedbacks and other product related textual descriptions (post redaction of confidential information), we instruct the text-only LLM to generate questions and answers as if it were looking at the image (while only text content was provided). Based on our observations, we discovered that employing a text-only LLMs for generating labels was adequate, as the general-domain multi-modal LLMs demonstrated suboptimal summarization capabilities when provided with both image and text inputs, likely due to limited capability of the model in understanding product images and inherent noise present in the sourced image-text pairs. Mostly, the textual data (submitted along with the image) from customer contacts tend to describe the products and their property/condition/issue from the customer's perspective, and hence can be used for formulating meaningful questions and answers. For catalog data, we only use the product information (title, description, etc.) as context. Using these contexts, we generate different types of instruction-following data encompassing diverse tasks for e-commerce. We also add few-shot examples to the prompt to illustrate the high-quality question-answer pairs for each task type based on the provided context. See Appendix A for the prompt template. Table 1 shows an example of instruction-following data. To mitigate data bias, we employed stratified sampling techniques. This was necessary because the original e-commerce data showed disproportionate representation of certain product categories, brands, and complaint types within specific timeframes. Our sampling approach ensured balanced representation across multiple dimensions including products, brands, geographical regions, customer issues, and time periods, resulting in a more comprehensive

and representative dataset. We finally collect 2M instruction-following samples in total, to represent diverse tasks and products.

3.2 Visual Attention Guided Data Refinement

The generated instruction-following dataset can be noisy due to fine-grained visual grounding errors, where certain segments of the textual descriptions may not be visually grounded. To alleviate this noise in the dataset, we need to analyze the visual grounding of the text with respect to the input image. There are several ways to check the *visual groundedness*, including semantic similarity based metrics such as CLIPScore (Hessel et al., 2022), SPICE (Anderson et al., 2016), etc., consensus based metrics (Vedantam et al., 2015) and attention visualization (Vig and Belinkov, 2019). Since attention maps offer a human-understandable measure of the weight given to the visual content during reading/generation of states, more so than other model internals, they provide a compelling signal for detecting visual grounding errors and more importantly, provide a fine-grained visual grounding information at token level. Formally, the attention mechanism is defined by the attention equation, which computes the attention scores between a query (Q) and key (K) :

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

Here, both K and Q are represented by concatenating the visual and text tokens and the self attention takes care of computing the dependency between the two modalities.

An aggregated visual attention score $A^V_{avg}(t)$ for the token at t is computed as the average attention weight across the V image tokens, L layers, and H heads:

$$A^V_{avg}(t) = \frac{1}{L \times H \times V} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{v=1}^{V} A_v^{(l,h)}, \quad (2)$$

where $A_v^{(l,h)}$ is the scalar attention weight of the token at t on the v^{th} image token in head h and layer l. Finally, for the text segment of interest, the above aggregated attention values are averaged across t to derive the overall score (VisAttnScore).

In practice, we can input the samples from the instruction-following dataset to any pre-trained VLM, compute the visual attention scores for the text tokens and eliminate samples with low visual grounding. Table 2 illustrates the relationship between visual attention score and visual grounding

Sample 1: (*CLIPScore=54%, VisAttnScore=56%*)
The image shows a shampoo bottle in leaking condition.

Sample 2: (*CLIPScore=54%, VisAttnScore=44%*)
The image shows a a leaking shampoo with contents spilled all over the carton box.

Sample 3: (*CLIPScore=52%, VisAttnScore=39%*)
The image shows a 100 ml shampoo, while the product description states a 200 ml conditioner.

Table 2: Illustration of how the visual attention score (colored as red, blue and green in increasing order of their magnitude) can be correlated to the visual grounding of text tokens.

for the text description. We clearly observe that the aggregated visual attention scores tend to be higher for visually grounded tokens and drop significantly for others. Alternatively, CLIPScore, with its focus on overall image-text similarity, is highly insensitive to fine-grained visual grounding errors between the feedbacks and is unsuitable to identify token-level visual grounding information. Our observation is consistent with the recent robustness study on image captioning evaluation metrics (Ahmadi and Agrawal, 2024). Additionally, we performed a pilot study with human annotators on a subset of 200 samples from the dataset to validate the reliability of these scores in identifying visually grounded texts. We observed 36% improvement in accuracy using our method in comparison to CLIPScore. In our experiments, we used pretrained IDEFICS2 for extracting visual attention scores and eliminated 25% of the dataset due to low visual grounding, resulting in 1.4M instruction-following samples of good quality. Furthermore, we filter out samples (≈5%) that contains images with less OCR or object detections and when the text is too short.

3.3 Adapting General Purpose Multi-Modal Model to E-Commerce Domain

To effectively adapt a general-purpose VLM to a new domain, the compute friendly method is to align and optimize only the vision-language connector module (while keeping the vision and language models frozen) on domain specific data. However, the information bottleneck in the frozen unimodal models requires domain concept feature

alignment on a high-quality large scale image captioning dataset, which is data intensive and not readily available in e-commerce domain. On the other hand, unfreezing and optimizing the full model (vision, language and vision-language connector modules) is highly compute intensive and requires several high-end GPUs. In contrast, we train with LoRA adapters (Hu et al., 2021) injected into all modules and find that this leads to faster, efficient and optimal domain adaptation with significantly lesser compute and data requirements. This serves as an efficient way for both concept alignment to e-commerce domain and impart instruction-following ability.

We use the IDEFICS2 (Laurençon et al., 2024) as our base VLM and continuously train the model for e-commerce domain with LoRA on our multi-task instruction-following dataset. IDEFICS2 employs Mistral-7B (Jiang et al., 2023) as the language model, SigLIP-SO400M (Zhai et al., 2023) as the vision encoder and a MLP projector with Perceiver Resampler (Jaegle et al., 2021) based pooling to connect the vision encoder and language model. It utilizes a fully-autoregressive architecture where the vision encoder's output is concatenated with text embeddings, and the entire sequence is fed into the language model optimized for next-token prediction loss. IDEFICS2 can process the images at their native resolutions and aspect ratio with NaViT strategy (Dehghani et al., 2023) and allows sub-image splitting (Li et al., 2024). For each sample, given the image (along with extracted OCR text) and instruction as input, we ask the model to predict the response and compute loss only on response tokens. We employ LoRA (r=256, α=32, dropout=0.1) applied to the attention layers of all transformer blocks. We fine-tune for 2 epochs with a initial learning rate of 2e-4 on 40 Nvidia A10G GPUs with a batch size of 8 per device. By removing noisy samples using the proposed filtering strategy (subsection 3.2), the total training duration reduced from 124 to 96 hours. We use AWS Textract for OCR extraction.

4 Experiments

4.1 Multi-Modal Benchmark for Product Images (MMPI-Bench)

Motivated by public VLM benchmarks like MM-Bench (Liu et al., 2024) and MME (Fu et al., 2023), we curated an internal e-commerce benchmark (MMPI-Bench) comprising a manually verified evaluation set of 6000 samples for three popular e-commerce tasks (equal samples), namely, Attribute Extraction (AE), Damage Detection (DD) and Product Matching (PM) from our test set, featuring products unseen during training. Our benchmark includes two types of evaluations using distinct instructions, *(i) MMPI-Obj-Bench*, measures objective (discriminative) capability via binary yes/no classification setup (balanced) and, *(ii) MMPI-Gen-Bench*, measures generative (visual reasoning) capability by leveraging an expert LLM (ClaudeV2) to evaluate the correctness of the model generated detailed answers with ground truth. Selected samples are presented in Appendix B and Appendix D.

Models	AE	DD	PM
InstructBLIP-14B	+2.1	+2.2	+2.4
Qwen-VL-9B	+7.4	+8.1	+7.4
IDEFICS2-8B	+8.8	+14.3	+17.7
IDEFICS2-8B (w/ ICL)	+11.2	+17.3	+20.6
ClaudeV3	+2.0	+14.3	+10.5
ClaudeV3 (w/ ICL)	+5.5	+18.9	+15.2
VIT-Pro (ours)	**+25.3**	**+23.8**	**+24.9**

Table 3: Quantitative evaluation on MMPI-Obj-Bench (relative to LLaVA-13B). AE: Attribute Extraction (only Brand), PM: Product Matching, DD: Damage Detection.

4.2 Main Results

Table 3 reports the accuracy-scores (relative to LLaVA-13B) of state-of-the-art multi-modal baselines and our instruction-tuned model (VIT-Pro) on MMPI-Obj-Bench. Among the generic VLMs, IDEFICS2 shows compelling performance in zero-shot setting with significant gains on DD and PM tasks. Further, when the baselines were evaluated in a few-shot setting with 4 examples each, we observed 5-10% performance increase with respect to their zero-shot evaluation results. VIT-Pro, reaps the benefit of visual-instruction tuning on domain-specific data, to achieve superior performance on all three tasks with a 11% improvement over IDEFICS2 and 15% gain over ClaudeV3 with in-context learning examples. For pretrained models, ICLs improved performance on average by 5-7%, but for our finetuned model we did not observe any notable gain. We tried two approaches for selecting ICL examples: manually curated examples and randomly selected examples matching the query's product category. Notably, carefully handpicked representative examples outperformed random sampling of examples, highlighting that the quality of ICLs can affect the performance

Models	AE	DD	PM
ClaudeV3	-8.8	-1.5	+38.2
VIT-Pro (ours)	**+30.6**	**+22.2**	**+43.2**

Table 4: Quantitative evaluation on MMPI-Gen-Bench (relative to IDEFICS2).

gains. We show additional results on AE task in Appendix C and qualitative analysis in Appendix D.

Table 4 reports the accuracy scores (relative to IDEFICS2-8B) on MMPI-Gen-Bench, calculated based on an expert LLM's decision. The LLM is prompted to provide a one-word answer, along with reasoning, on whether the ground truth matches the predicted detailed answer. If the LLM's decision is "Yes", it implies the ground truth answer matches the predicted answer. We observe a significant accuracy drop compared to the discriminative task metrics in Table 3. This clearly indicates that while the models are proficient at providing objective answers, they need improvement in detailed reasoning, providing actual facts, and reducing hallucinations.

4.3 Ablation Studies

We conducted detailed ablation experiments and robustness studies to understand the VIT-Pro's limitations under different settings. Specifically, it includes several robustness tests with respect to additional inputs (OCR, images), image resolution/splitting, LoRA adapters and effect on using model optimization strategies like 4-bit quantisation, flash-attention, etc. The key results from this series of ablation are captured in Table 5, Table 6 , Table 7, Table 8 and the remaining are discussed in Appendix.

OCR. Removing OCR from the inference prompts significantly degraded performance across most tasks. PM task saw the most substantial degradation, as OCR helps in extraction of fine-grained textual details from images. However, DD task relies solely on visual cues rather than textual information in product images, and AE task, esp. for brand can be easily handled without OCR.

Resolution. The average image resolution in the MMPI-Bench is around 1400×1200 pixels. While VIT-Pro was trained with native resolution (up to 980x980) and native aspect ratio, we tested four input resolutions during inference: native, 224x224, 512x512, and 768x768. As shown in Table 5, resizing to 224x224 impairs performance, with DD (which solely relies on visual tokens) exhibiting the most significant degradation. Tasks like AE and PM may still benefit from OCR. However, we observe diminishing returns beyond 512x512 resolution. This suggests that while customer-clicked images from modern smartphones may have high resolution, resized 512x512 images should suffice for similar e-commerce vision tasks.

Image Splitting (IS). Image splitting enables passing images of very large resolution by dividing each input image into 4 sub-images and concatenating them with the resized original to form 5 images. Disabling image splitting led to a slight decrease in the model's performance, but improved model latency.

Multi-image. Customer issues often involve multiple images captured from different touchpoints, offering unique perspectives and details. We conducted an ablation study on VIT-pro's performance with single and multiple images for DD and PM tasks, which require cross-image correlations grounded in both visual and textual information, challenging for traditional VLMs. For PM, one reference image was provided as context for visual comparison, while for DD, 2-3 item perspectives were given to assess condition. As shown in Table 5, multi-image training significantly improved performance on both tasks, increasing PM accuracy scores by +6%, and DD by +0.4%. We discuss the training details and samples in Appendices Appendix E.

LoRA adapters. We ablate the usage of LoRA adapters in the different model components and show the resulting performance effect on MMPI-Bench in Table 6. Interestingly, against common practice of keeping the language model (*LM*) layers frozen, we notice that LoRA based learning is most critical in the LM layers. Freezing the LM layer results in a significant drop (-11.1%) in overall performance driven majorly by *Product Matching* and *Attribute Extraction* - tasks where comprehension of the language aspects are critical for performance. We attribute this observation to the lack of sufficient domain knowledge with the LLMs on product label related text and linguistics. Similarly, following standard practice of fine-tuning only the modality connector module (*VLC*) is insufficient and results in a large drop (-21%) of performance. Finally, freezing only the vision encoder (*VM*) results in the least drop (-8.2%) in model performance in-

OCR	IS	Resolution				Multi-Image	AE	DD	PM	Latency (sec/it)
		Native	224x224	512x512	768x768					
✓	✓	✓					*	*	*	*
	✓	✓					-1.0	-0.9	-9.0	-0.3
✓		✓					-1.8	-2.4	-0.6	-0.5
✓	✓		✓				-0.8	-21.9	+0.4	-0.3
✓	✓			✓			0.0	-0.9	+1.5	-0.2
✓	✓				✓		0.0	0.0	+0.6	-0.1
✓	✓	✓				✓	-	+0.4	+6.0	+1.1

Table 5: Ablation studies under different settings using VIT-Pro on MMPI-Obj-Bench. The quantitative numbers reported are relative to the default setting in first row.

dicating the generalisability of the SigLIP vision model on in-the-wild product images.

VLM Components			AE	DD	PM	Overall
VM	VLC	LM				
✓	✓	✓	*	*	*	*
✓			-22.5	-23.1	-22.2	-22.6
	✓		-19.8	-21.6	-21.9	-21.0
		✓	-26.1	-16.7	-19.7	-20.8
	✓	✓	-7.4	-7.3	-10.0	-8.2
✓		✓	-22.1	-22.3	-22.2	-22.1
✓	✓		-15.4	-6.3	-11.7	-11.1

Table 6: Effect of applying LoRA adapters across the 3 submodules: Vision Model (*VM*), Vision Language Connector (*VLC*) & Language Model (*LM*) for VIT-Pro. We report the performance numbers relative to the default setting in first row.

Impact of Quantization & Flash Attention. We further ablate the use of 4-bit quantization and Flash Attention 2 in VIT-Pro. Table 7 illustrates the impact of 4-bit quantization on model performance, latency, and memory usage. While quantization significantly reduces memory consumption by 11.8 GB and improves latency by 0.1 secs/it, it comes at a slight cost to performance across AE, DD, and PM tasks. Table 8 demonstrates the effects of using Flash Attention 2, showing marginal improvements in task performance (AE: +0.1, DD: +0.1, PM: +0.4) while substantially reducing latency by 0.45 secs/it. For high-throughput, real-time e-commerce applications, these substantial improvements in memory usage and latency are crucial. Despite the slight reduction in model performance, the 4-bit-quantized version with Flash Attention 2 emerges as the preferred implementation choice. The significant gains in efficiency and speed make it particularly well-suited for e-commerce operations, where rapid response times and resource optimization are paramount.

Quant (4-bit)	AE	DD	PM	Latency (sec/it)	Memory (GB)
✗	*	*	*	*	*
✓	-1.4	-1.8	-0.9	-0.1	-11.8

Table 7: Impact of quantization on model performance, latency and memory usage.

FlashAttention2	AE	DD	PM	Latency (sec/it)
✗	*	*	*	*
✓	+0.1	+0.1	+0.4	-0.45

Table 8: Impact of Flash Attention 2 on model performance and latency.

5 Conclusions

We showcased the potential of leveraging large-scale weakly-associated image-text pairs commonly available in any e-commerce stores to build a multi-task vision-language model for e-commerce domain. VIT-Pro, demonstrates superior performance over open-source and commercial baselines on an internal e-commerce vision-language benchmark. Comprehensive analyses highlight VIT-Pro's robustness under varying input configurations like resolutions, OCR, multi-image scenarios, optimization strategies and LoRA adapters. In future, we want to incorporate other data sources (e.g. X-Rays) and tasks (e.g. product grading).

References

Saba Ahmadi and Aishwarya Agrawal. 2024. An examination of the robustness of reference-free image captioning evaluation metrics. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 196–208, St. Julian's, Malta. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson,

and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. *Preprint*, arXiv:1607.08822.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *Preprint*, arXiv:2308.06595.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *Preprint*, arXiv:2311.12793.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey Gritsenko, Mario Lučić, and Neil Houlsby. 2023. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. *Preprint*, arXiv:2307.06304.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*.

Jinmiao Fu, Shaoyuan Xu, Huidong Liu, Yang Liu, Ning Xie, Chien-Chih Wang, Jia Liu, Yi Sun, and Bryan Wang. 2022. Cma-clip: Cross-modality attention clip for text-image classification. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2846–2850.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. *Preprint*, arXiv:2103.03206.

Qinjin Jia, Yang Liu, Daoping Wu, Shaoyuan Xu, Huidong Liu, Jinmiao Fu, Roland Vollgraf, and Bryan Wang. 2023. KG-FLIP: Knowledge-guided fashion-domain language-image pre-training for E-commerce. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 81–88, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Anant Khandelwal, Happy Mittal, Shreyas Kulkarni, and Deepak Gupta. 2023. Large scale generative multimodal attribute extraction for E-commerce attributes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 305–312, Toronto, Canada. Association for Computational Linguistics.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. *Preprint*, arXiv:2311.06607.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv:2304.08485*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

OpenAI. 2023. Gpt-4 technical report.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *Preprint*, arXiv:1411.5726.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *Preprint*, arXiv:1906.04284.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

A Prompts

Table 9 shows the prompt template used for producing visual instruction-following data.

B Samples from MMPI-Bench

Samples from MMPI-Obj-Bench are shown in Table 10 and qualitative samples from MMPI-Gen-Bench are shown in Table 12.

C Attribute Extraction Performance

We present the attribute extraction performance on all key attributes in Table 11 using images from MMPI-Bench. The task involves generating a JSON object with attribute names and values extracted from the input image (refer to the example prompt in Table 1). We employ an exact string match after normalizing the ground truth and predicted string values. We notice that across majority of the attributes, VIT-Pro achieves a significant performance gain compared to ClaudeV3 and IDEFICS2.

D Qualitative Results

Table 12 illustrates the qualitative performance of models using task-specific instructions to study their generative capability. General-domain VLMs exhibit limited zero-shot capabilities for domain-specific use cases. Their sub-optimal performance can be attributed to: *a)* limited effectiveness on in-the-wild images with partially visible regions, occlusions & poor-quality, and *b)* limited generalization to out-of-domain and complex visual reasoning tasks. VIT-Pro bridges this domain gap, showing promising visual recognition and reasoning capabilities for the e-commerce domain.

E Multi-Image Reasoning

Training Setup. We curate a multi-image version of our instruction-following dataset in a similar fashion, with number of images ranging from 2-5 per task. For each sample, the model predicts a response based on the input images (including OCR text) and instruction and the loss is calculated exclusively on the response tokens. We employ LoRA with a much lower rank of r=8, a scaling factor of α=16, and a dropout rate of 0.1 applied to the attention layers of all transformer blocks. Model is fine-tuned for 2 epochs with a lower initial learning rate of 1e-5 on 8 Nvidia A10G GPUs with batch size of 16 and gradient accumulation steps of 8. Through careful hyperparameter selection and controlled parameter adaptation through LoRA, we improve training stability on our multi-image dataset.

Prompts. We observe that fine-tuning VLMs, that are largely pre-trained over single-image datasets, with the multi-image complexity is highly sensitive to prompt structure especially with multiple images as context. Adding delimiters like ###, <<< >>> specify the boundary between different sections of the prompt. We follow a numbering style for images in our prompts instead of stacking images together. This creates a distinct image separation for LLM's multi-image reasoning. Table 13 shows the formatted prompts for PM and DD tasks, suitable for the multi-image visual comparison and visual reasoning tasks. In our example, we used ### to indicate difference in contexts and numbers like [1], [2] in front of images to indicate clear distinction in the contexts and images. We observe that this makes the VLM's output less sensitive to the changes in image ordering.

Qualitative Samples. Samples from multi-image version of MMPI-Obj-Bench dataset as shown in Table 13 demonstrate the complexity in the multi-image reasoning. For non-trivial scenarios, customers share multiple product images, either a) multiple views to better articulate the item state or b) multiple perspectives as supporting evidences to strengthen their claims. First two example shows a scenario where the customer highlights that the leakage from ghee jar from different views and its soiled packaging. Here, the multi-image VLM capability that performs visual comparison, co-reference and reasoning across images is needed for a confident assessment. We further see the usefulness of having an additional images to improve the models decision making. Third example shows a scenario where the supplied image appropriately matches the product description however,

Prompt template to generate visual instruction-following data

```
User: You are an AI assistant well-versed in e-commerce product images. You are provided with a
context in the form of customer feedbacks/chats and possibly additional context about an e-commerce
product image. Unfortunately, you don't have access to the actual image. Design questions and answers
about the product, as if you are seeing the image.

Rules for generating question and answer pairs:
1) Ask diverse questions and visually grounded answers.
2) Questions should be about the visual content of the image, including product type, counts,
attributes, condition, package, positions, product comparison, etc.
3) For questions that do not have a definite answer given the limited context, acknowledge it and
politely refuse to answer with valid reasons.
4) Include questions that requires different response formats like list, json, short text,
detailed text, etc.
(....remaining rules omitted for brevity)

Context related to customer feedback:
<context_1>{CONTEXT_1}</context_1>

Context related to product information:
<context_2>{CONTEXT_2}</context_2>

Here are a few examples:
<examples>
    <example>
        <context_1>...</context_1>
        <context_2>...</context_2>
        <question></question>
        <answer></answer>
    <example>
    (....remaining examples omitted for brevity)
</examples>

Assistant:
```

Table 9: Prompt template to generate visual instruction-following data

the visual comparison with the reference image adequately helps with the decision making. Fourth example show cases a scenario where the different views of the image are used to retrieve relevant information such as product brand and item weight. We see that in e-commerce tasks where textual descriptions alone are not sufficient, addition of a reference image enriches the context for holistic decision making.

F Industry Impact

Currently, manual investigations form the backbone of resolving multi-modal queries, such as those involving quality and quantity assurance of the delivered product. Auditors manually examine captured images alongside textual descriptions to verify issues like packaging errors, delivery time damages, product quality, etc. However, this approach is neither scalable nor efficient for the massive scale of modern e-commerce operations. To automate investigations, the proposed VIT-Pro could be directly leveraged. To evaluate the potential real-world impact, we conducted a 4-week shadow mode experiment in co-pilot setup across three tasks: damage detection, product matching, and attribute extraction. Results showed significant improvement in investigation efficiency and decision quality thereby enhancing customer experience through faster and more precise decisions. VIT-Pro can seamlessly integrate into other applications in e-commerce stores requiring multi-modal understanding to scale operations. We strictly adhered to ACL code of ethics and professional conduct during the course of this research (refer Appendix G).

G Ethics Statement

We used e-commerce data from customer refund/return claims and product catalogs, with consent. We carefully redacted any personally identifiable information from the data, preventing any misuse /adverse impact. Our data curation strategy requires no

Task	Prompt Image	Prompt Text	Label
Damage Detection		**Instruction:** provide an answer to the question in a single word. Use the image to answer. **Question:** Is there a damage on the product in the image? OCR Tokens: <ocr> **Answer:**	No
		**Instruction:** provide an answer to the question in a single word. Use the image to answer. **Question:** Is there a damage on the product in the image? OCR Tokens: <ocr> **Answer:**	Yes
Product Matching		**Instruction:** provide an answer to the question in a single word. **Product Description:** Set Wet Hair Wax For Men - Fibre Hair Wax 60g | Strong Hold, Extra Volume, Natural Finish, Restylable Anytime, Easy Wash Off **Question:** Does the image match to the product description? OCR Tokens: <ocr> **Answer:**	No
		**Instruction:** provide an answer to the question in a single word. **Product Description:** Cross 8910-1 Slim RollerBall Gel InkRefill for Spire,Click and Classic Century, Black **Question:** Does the product description match with the image? OCR Tokens: <ocr> **Answer:**	Yes
Attribute Extraction		**Instruction:** provide an answer to the question in a single word. Use the image to answer. **Question:** Is this product from "DOVE" brand ? OCR Tokens: <ocr> **Answer:**	No
		**Question:** Is this product from "Del Monte" brand? OCR Tokens: <ocr> **Answer yes or no:**	Yes

Table 10: Samples from MMPI-Obj-Bench

Attributes	IDEFICS2	ClaudeV3	VIT-Pro
ProductType	*	+2.7	**+61.2**
Brand	*	**+10.6**	+2.3
Size	*	-4.9	**+17.0**
Count	*	-26.0	**+10.8**
TargetGender	*	-6.5	**+45.7**
Color	*	-12.5	**+19.4**
ExpiryDate	*	+2.9	**+9.1**

Table 11: Attribute-level performance comparison (relative to IDEFICS2).

human annotation, except for manually verifying the MMPI bench evaluation set. For this, annotators are compensated following industry standards. The impact estimation of VIT-Pro through shadow mode experiment was conducted in co-pilot mode, with human experts vetting model decisions to mitigate risks. Our research maintains objectivity and adheres to the ACL code of ethics and professional conduct.

Qualitative Examples from MMPI-Gen-Bench	
	**User:** Describe the image and comment on the product condition. **ClaudeV3:** The image shows a person's hand holding a bottle of Persil ProClean laundry detergent. The bottle appears to be intact and in usable condition, though slightly crumpled or dented at the bottom portion. **IDEFICS2:** The product in the picture is a bottle of hand sanitizer. It is not damaged. **VIT-Pro:** The image shows a damaged product packaging with the lotion leaking out. The product seems to have spilled out of the bottle.
	**User:** Does the description match with the product in the image? Provide reasons. Product Description: Nivea Extra Whitening Cell Repair Body Lotion, Spf 15, 400ml. **ClaudeV3:** No, the product description does not match the product shown in the image. Reasons: 1. The product in the image is labeled as "Nivea Natural Glow Body Lotion with Cell Repair", while the description mentions "Nivea Extra Whitening Cell Repair Body Lotion". 2. The image does not show any mention of SPF 15, whereas the description includes "SPF 15". 3. The net quantity mentioned in the description is 400ml, but the quantity is not visible on the product in the image. **IDEFICS2:** Yes, the description matches with the product in the image. The product is a body lotion with SPF 15 and is 400ml in size. The Nivea logo is visible on the bottle, and the label on the bottle also mentions that it is an extra whitening cell repair body lotion. The bottle is brown, which is consistent with the description. **VIT-Pro:** No, the product description does not match the image. The image shows a bottle of Nivea Natural Glow Body Lotion with Cell Repair and Vitamin C. The product name on the bottle mentions "Natural Glow" and not "Extra Whitening" as stated in the given product description.
	**User:** Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put N/A instead). **Format:** {"Product Name": xxx, "Brand": xxx, "Count": xxx, "Color": xxx, "Target Gender": xxx, "Size": xxx, 'Manufacture Date': xxx , 'Expiry Date': xxx } **ClaudeV3:** { "Product Name": "Vitamin E++ Skin Oil", "Brand": "Richfeel", "Count": "1", "Color": "N/A", "Target Gender": "N/A", "Size": "500ml", "Manufacture Date": N/A, "Expiry Date": "02/23" } **IDEFICS2:** {"Product Name": "vitamin E++ skin oil", "Brand": "Richfeel", " Count": "NA", "Color": "NA", "Target Gender": "NA", "Size": "NA", "Manufacture Date": "2022-01-01", "Expiry Date": "2023-01-01"} **VIT-Pro:** {"Product Name": "Richfeel Vitamin E++ Skin Oil, 500ml", "Brand": "Richfeel", "Count": 1.0, "Color": "White", "Target Gender": "unisex", "Size": "500 ml (Pack of 1)", "Manufacture Date": "02/2022", "Expiry Date": "01/2023"}

Table 12: Challenging examples from MMPI-Gen-Bench to compare visual reasoning and information extraction capabilities.

Task	Prompt Images	Prompt Text	Label
Damage Detection	Customer images 	**Instruction:** provide an answer to the question in a single word. Use the image to answer. [1] <*image1*> [2] <*image2*> [3] <*image3*> [4] <*image4*> **Question:** Is there a damage on the product shown in the images? OCR Tokens: <ocr> **Answer:**	Yes
		**Instruction:** provide an answer to the question in a single word. Use the image to answer. [1] <*image1*> [2] <*image2*> [3] <*image3*> **Question:** Is there a damage on the product shown in the images? OCR Tokens: <ocr> **Answer:**	Yes
Product Matching	Customer Image Reference Image 	**Instruction:** provide an answer to the question in a single word. ### Customer shared images: [1] <*image1*> ### OCR Tokens from Customer shared images: <*ocr*> ### Reference Product's Image: <*ref-image*> **Product Description:** [BRAND] Navy Blue Colour with Yellow Stripes Design Calf Length School Cotton Socks for Boys & Girls (Pack of 5 Pairs) **Question:** Do the customer submitted images match the product? Use the product's description and image to answer. **Answer:**	No
		**Instruction:** provide an answer to the question in a single word. ### Customer images: [1] <*image1*> [2] <*image2*> ### OCR Tokens from Customer shared images: <*ocr*> ### Reference Product's Image: <*ref-image*> **Product Description:** [BRAND] A2 Bilona Desi Cow Ghee 500 ml - Pure Brijwasi Ghee - Bilona Curd Churned - Lab Tested - Perfect Aroma & Danedar Ghee - Grass Fed **Question:** Do the customer submitted images match the product? Use the product's description and image to answer. **Answer:**	Yes

Table 13: Here are few samples from multi-image version of MMPI-Obj-Bench that demonstrate the complexity in the multi-image reasoning.