Adapting Vision-Language Models for E-commerce Understanding at Scale

Matteo Nulli[1,2,*], Vladimir Orshulevich[1], Tala Bazazo[1], Christian Herold[1], Michael Kozielski[1], Marcin Mazur[1], Szymon Tuzel[1], Cees G. M. Snoek[2], Seyyed Hadi Hashemi[1], Omar Javed[1], Yannick Versley[1] and Shahram Khadivi[1]
[1]eBay Inc., [2]University of Amsterdam
{mnulli, tbazazo}@ebay.com

Abstract

E-commerce product understanding demands by nature, strong multimodal comprehension from text, images, and structured attributes. General-purpose Vision–Language Models (VLMs) enable generalizable multimodal latent modeling, yet there is no documented, well-known strategy for adapting them to the attribute-centric, multi-image, and noisy nature of e-commerce data, without sacrificing general performance. In this work, we show through a large-scale experimental study, how targeted adaptation of general VLMs can substantially improve e-commerce performance while preserving broad multimodal capabilities. Furthermore, we propose a novel extensive evaluation suite covering deep product understanding, strict instruction following, and dynamic attribute extraction.

1 Introduction

Deep e-commerce product understanding is inherently multimodal. While today's search works primarily through matching the textual part of a listing, images of an item, its packaging, or general visuals play a large role in how customers evaluate and select the item they want. Recent advancements in Large Language Models (LLMs) (Dubey et al., 2024; Yang et al., 2024; Mistral AI, 2024), have shown strong results on e-commerce tasks, with some specific approaches for domain-specific customization (Peng et al., 2024; Herold et al., 2025). However, translating these gains into the vision–language setting, like we do in this paper, remains a considerable challenge.

With the advent of general-purpose Vision–Language Models (VLMs) such as LLaVA-OneVision (Li et al., 2024a), Qwen3-VL (QwenTeam, 2025), InternVL3 (OpenGVLab-Team, 2024), and Gemma3 (Gemma-Team, 2025), deploying multimodal systems in e-commerce has become feasible. Nevertheless, we see a need for a reproducible, backbone-agnostic recipe for adapting VLMs to the demands of e-commerce attribute-centric reasoning, multi-image aggregation, and robustness to noisy seller-generated content, *without* loosing general VLM-capabilities performance. Moreover, in spite of a large amount of evaluation sets for text-only shopping tasks (Jin et al., 2024), rigorous benchmarking of multimodal shopping assistants remains underdeveloped.

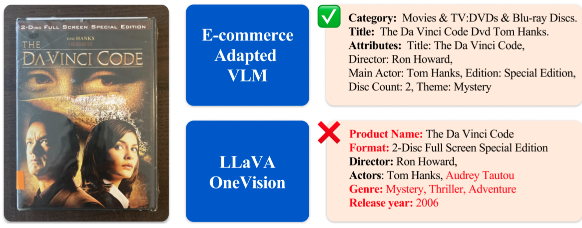

Figure 1: **Output of our E-commerce Adapted VLMs compared against same size LLaVA-OneVision.** We show our models ability to more faithfully extract attributes from e-commerce items. In red, we highlight wrong model predictions that are neither tied to the image nor valid item attributes.

In this paper, we focus on two questions, (i) if high-performing e-commerce VLMs truly require a customized LLM, or whether adapting on vision-focused tasks suffices. And (ii) on the best way to build a benchmark to assess multiple dimensions of understanding from extracting product attributes to category-specific deeper understanding and handling of multi-image tasks. To tackle (i) we perform **extensive ablations across multiple visual and text decoders** as backbones. Moreover we propose a new set of **multimodal instruction data** to strengthen e-commerce abilities without hindering general performance, showing adaptation is possible. To answer (ii), we propose a set of benchmarks evaluating a broad range of **internal use-cases and real-life online retail** scenarios. In summary our contributions are as follows:

- We show how to **adapt existing VLMs to-**

wards the e-commerce domain, taking into account task-specific features, and demonstrate it enhances performance on online shopping tasks considerably, without any loss of capabilities on other domains.

- We design and implement a comprehensive set of vision, **e-commerce benchmarks** based on real production problem statements and data.
- We also evaluate state-of-the-art VLMs across general-domain and in-domain multimodal tasks, reporting our adaptation findings across data mixtures, models sizes and architectures.

All in all we provide insights, evaluation suites and a proven strategy for an e-commerce adaptation of VLMs, retaining strong general capabilities.

2 Related Work

e-Commerce Vision Language Models Online shopping platforms own an enormous quantity of data which can be leveraged to train LLMs and VLMs. Among the many applications, the ability of models to concretely grasp user-uploaded *visual*-information, correctly comprehending multimodal product characteristics and being able to predict them accordingly are vital features in online marketplace applications. Research efforts such as Xue et al. (2024); Li et al. (2024b), finetune VLMs for product understanding and tackle product description generation exploiting in-context learning capabilities. Similar e-commerce adaptation works like Ling et al. (2024) instruction tune Llama-3.2 model with online shopping data. While these are interesting research directions, none have yet concurrently studied the effect of multiple pre-trained multimodal architectures on downstream online retail performance, all while being able to retain effectiveness on general purpose multimodal benchmarks.

E-commerce-specific Evaluation Text-centric suites (Jin et al., 2024) have helped standardize measurement of general shop-assistant abilities and even powered community competitions, but they operate primarily on textual signals. Similar widely used datasets evaluate query–product relevance, review-grounded product Q&A, purchase-intention comprehension and domain factuality via knowledge graphs (Reddy et al., 2022; Gupta et al., 2019; Ding et al., 2024; Chen et al., 2025a; Liu et al., 2025). While general-purpose VLM evaluations (Fu et al., 2024) stress broader visual-language understanding, like visual-question answering or object recognition, they are not tailored to the e-commerce fine-grained attributes and tool use typical of retail. In recent research, Ling et al. (2025) covers some question answering, product classification and relevance-related tasks as well as product relation identification and sentiment analysis and their dataset, while large-scale and comprehensive, is built by taking text-only datasets, adding images and removing the image-text pairs where the images are redundant, whereas we feel that our setting of taking image-focused tasks as a starting point is more naturalistic.

3 Methodology

3.1 Our E-commerce Benchmarks

To tackle the gap in multimodal e-commerce-specific benchmarks, we propose a set of four evaluation suites described below. Each is designed to tailor internal production use-cases, ranging on a variety of tasks, categories and metrics.

Aspect Prediction Our Aspect Prediction evaluation set, divided into three different sub-parts. The first, comprised of 2600 general questions on all e-commerce categories, and the second two, evaluate the model's ability to predict aspects in Fashion, with and without additional contexts from item title and category, both with 1600 examples. All are evaluated through string matching.

Deep Fashion Understanding We design a specialized benchmark consisting of 3000 samples divided into three subsets: *Apparel Men Shirts and Women Tops*, *Handbags*, and *Sneakers*. Each subset targets critical attributes relevant to the product type, structured into clear classification categories. Evaluation involves prompting the model to categorize items precisely according to the provided attribute classes.

Dynamic Attribute Extraction This evaluation set comprises 1,000 synthetically generated with GPT-4o (gpt, 2024), human-verified examples. It benchmarks a model's ability to enumerate and structure all visually grounded attributes from an image without a predefined schema.

Multi-image Item Intelligence In this dataset the model is asked to compile a fixed set of attributes related to compliance questions (e.g. brand,

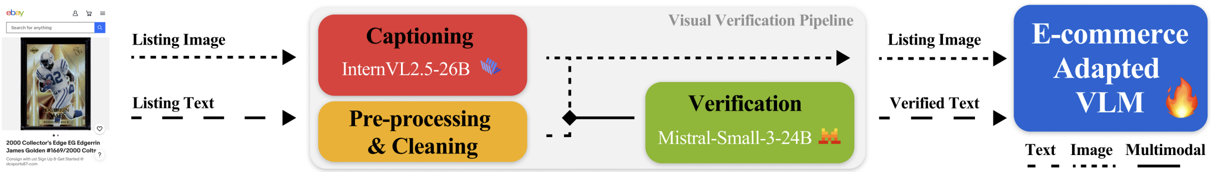

Figure 2: **Visual Verification Pipeline.** The figure shows the pipeline we use to create the 4M e-commerce visual instruction tuning data. We begin by collecting raw listings data from the web (*left*). We then clean and pre-process the textual entries. In parallel, we create detailed *captions* for the corresponding image through InternVL-2.5-26B. Finally, we provide the *captions* together with the *cleaned listings* to Mistral-Small-3-24B to obtain the *verified* instructions, used, along with original images, to train our models (shown with fire).

warning labels, ingredients) from multiple product items into a structured `JSON` output, enabling verification and recall matching processes. 1000 items were sampled to prioritize product categories with high regulatory requirements (toys, electronics, electrical appliances, cosmetics, etc.). We evaluate through LLM-as-a-judge (see e.g. Gu et al., 2025). More on each set in Appendix A.4.

3.2 Our Approach to E-commerce Adaptation

We first go through our Data Curation pipeline, VLM Adaptation Training Stages, additional Multi-Image Item Intelligence specific fine-tuning and the architectures on which we apply this adaptation.

3.2.1 Internal Data Curation

Raw e-commerce listings data is typically rather noisy, yet high-quality data is crucial when training large multimodal models. Here, we show how to leverage the self-supervised signal inherent in user-generated listings data and describe our *Visual Verification Pipeline* for large-scale data curation, illustrated in Figure 2. We begin by collecting nearly 15 million raw listings from online marketplace websites and select only the primary (main) image for each listing. Each image is captioned through InternVL-2.5-26B (Chen et al., 2025b). Alongside, we extract the user-supplied item aspects from each listing. Given the generated caption and item aspects, we employ Mistral-Small-3-24B (Mistral AI, 2024) to verify which of these aspects can be inferred from the caption and thus from the image itself. This verification ensures visual-textual correspondence during training. The resulting listings, enriched with the verified aspects and paired with their original images, form the high-quality dataset used to train our multimodal models.

3.2.2 General E-commerce Adaptation

Following LLaVA-OneVision (Li et al., 2024a), we train our models in three stages: (i) Vision-Language Alignment, (ii) Mid-Stage Training, and (iii) Visual Instruction Tuning. For (i) we employ LLaVA-OneVision set of instructions with BLIP-LAION 558k corpus (Liu et al., 2023) and for (ii) their mid-stage mixture (Li et al., 2024a) removing several subsets that we found low-signal or redundant.

Visual Instruction Tuning Finally, we conduct instruction tuning on (a) a version of the LLaVA-OneVision single-image mixture, and (b) ∼4M internal e-commerce oriented set of instructions pictured in Appendix Figure 4. This portion is partitioned as follows, with percentages equaling part of e-commerce total: **VQA** (45%), consists of free-form, yes/no, image-only questions, full item description all with and without title & category context. **Dynamic Attribute Extraction** (30%), containing free-form visual attribute extraction with and without title & category context. Variants include augmenting it with OCR, prompt constraining text, and any combinations of these settings. **Precise Instruction Following** (12.5%), a set of keyword-conditioned instructions that require inclusion/avoidance of specific terms and tasks emphasizing strict form/length control. **Listings** (12.5%), comprised of full product listings predictions from an image. Details in Appendix A.7.

3.2.3 Item Intelligence Fine-Tuning

For the Multi-Image Item Intelligence task, we curate a fine-tuning dataset of 100,000 items across relevant categories, each containing multiple images (median = 5, range = 2–8). Since no labeled data is available, we generate first annotations using GPT-4.1 via prompt-engineering. We then enhance annotation quality, by refining supervision to focus on visually and semantically informative regions—often textual or numeric details on product surfaces. We achieve this employing Qwen2.5-VL-32B (Bai et al., 2025) to produce precise bounding boxes, which are post-processed (expanded and

Vision Encoder ǀ LLM		Aspect Prediction			Deep Fashion Understanding		Dynamic Attribute Extraction
		General	Fashion	Fashion + T&C	Apparel	Sneakers & Handbags	DAE
Internal E-commerce Adaptation							
[1] SigLIP2 ǀ Llama-3.1-8B		37.7	46.0	51.9	67.0	75.1	59.7
[2] SigLIP2 ǀ e-Llama3.1-8B		44.4	52.8	60.4	78.9	79.5	66.1
[3] Qwen2.5ViT ǀ e-Llama3.1-8B		53.3	55.1	65.3	71.0	70.1	**70.7**
[4] SigLIP2 ǀ Qwen-3-4B		54.6	**60.7**	67.5	78.6	80.1	66.5
[5] SigLIP2 ǀ Qwen-3-8B		**56.2**	60.1	**68.5**	**79.8**	**81.6**	68.1
[6] SigLIP2 ǀ Lilium-1B		41.0	48.4	54.4	72.2	71.0	66.3
[7] SigLIP2 ǀ Lilium-4B		42.3	49.1	56.7	74.7	73.5	68.3
[8] SigLIP2 ǀ Lilium-8B		42.4	49.2	55.8	75.2	77.0	68.0
[9] SigLIP ǀ Gemma3-4B		54.8	58.3	67.0	78.6	80.3	67.6
Open Source							
[10] SigLIP ǀ Qwen2-7B	*LLaVA-OV*	28.7	30.3	47.4	62.8	39.5	67.0
[11] Qwen2.5ViT ǀ Qwen2-7B	*Qwen2.5-VL*	36.9	36.8	47.7	82.9	80.6	72.0
[12] Qwen3ViT ǀ Qwen3-8B	*Qwen3-VL*	**40.5**	**42.4**	**58.2**	**84.3**	**84.6**	**70.9**
[13] SigLIP ǀ Gemma3-4B	*Gemma3*	24.3	29.0	40.4	64.2	77.5	72.7

Table 1: **Internal tasks comparison across model architectures and sizes.** We report performance of vision encoder and LLM combinations on three of our proposed evaluation sets (top row). Internal E-commerce Adaptation models indicate VLMs fully trained top to bottom starting from pre-trained backbones, Open Source indicates models not trained by us, the original *model names* are next to their architectural structure. Higher is better (↑).

merged) for better coverage. Cropped regions and original images are then re-annotated by GPT-4.1, yielding substantially higher-quality *better labels*. More details in Appendix A.5.

3.2.4 Model Architectures

We compare several state-of-the-art (SOTA) model components for our e-commerce VLM. For the vision encoder, we experiment with **SigLIP2-SO400M-Patch14-384** (Tschannen et al., 2025) and **Qwen2.5 ViT** (Bai et al., 2025). As text decoder, we compare **Llama3.1-8B** (Touvron et al., 2023), **e-Llama3.1-8B** (Herold et al., 2025) an e-commerce adapted version of Llama3.1 8B, **Lilium 1B/4B/8B** (Herold et al., 2024) trained from scratch for the e-commerce domain and **Qwen3 4B/8B** (Yang et al., 2025). Furthermore, we also adapt fully fledged SOTA VLMs for certain tasks, namely **Llama-3.1-Nemotron-Nano-VL-8B-V1**, **Gemma3 4B/12B/27B** (Gemma-Team, 2025), **Qwen2.5VL-7B** (Bai et al., 2025) and **Qwen3VL-8B** (QwenTeam, 2025).

4 Experiments

In our Experiments section, we compare our e-commerce adapted VLMs against existing ones (Section 4.2), followed by an analysis of the importance of vision encoders (Section 4.3) and text decoders (Section 4.4). In the second part, we focus on the item intelligence use-case (Section 4.5).

4.1 Experimental Setup

All models that we trained are optimized as described in Section 3.2. For training, we use the NeMo (Kuchaiev et al., 2019) and LLaVA-OneVision frameworks (Li et al., 2024a), using the same loss objective. Training was conducted on NVIDIA H100 GPUs (using up to 120 GPUs connected via NVLink and InfiniBand). In addition to our set of e-commerce benchmarks (see Section 3.1), we also evaluate all models on a comprehensive set of public benchmarks. For details, see Appendix A.2.

4.2 Comparison against existing VLMs

We first compare our initial internally trained VLM **SigLIP2 ǀ Llama-3.1-8B** against external VLMs as shown in Table 2 row 14 for general-domain benchmarks and in Table 1 row 1 for e-commerce tasks. We find that newer SOTA external VLMs like **Qwen3-VL-8B** outperform our internal model on the majority general-domain benchmarks. However, on the e-commerce specific benchmarks, the picture is quite different. While some external models do perform very well on Deep Fashion Understanding, they do fall behind on most e-commerce specific benchmarks. This leads us to the conclusion that we need to invest in building our own customized VLM for relevant e-commerce tasks. In the following sections, we determine the best overall settings to accomplish this goal.

4.3 Importance of Vision Encoder

We begin this exploration by analyzing the importance of the vision encoder, comparing two architectures, **SigLIP2** and **Qwen2.5 ViT** while keeping the text encoder the same. On both e-commerce tasks (compare Table 1 rows 2 & 3), and general domain benchmarks (compare Table 2 rows 15 &

Vision Encoder | LLM	Multimodal General Understanding				Vision	OCR, Chat/Doc QA		Reasoning	e-Commerce
	MMBench (dev)	MME (Perc.)	MME (Cogn.)	MMStar	CVBench	TextVQA (val)	AI2D (val)	MMMU (val)	eComMMMU (test)
Internal E-commerce Adaptation									
14 SigLIP2 | Llama-3.1-8B	75.8	1556.1	314.6	49.5	62.3	75.2	76.3	43.9	NA
15 SigLIP2 | e-Llama3.1-8B	76.9	1549.1	379.3	52.6	72.7	74.0	78.2	42.0	46.1
16 Qwen2.5ViT | e-Llama3.1-8B	71.7	905.8	333.2	53.6	61.6	65.2	76.6	39.7	45.9
17 SigLIP2 | Qwen-3-4B	81.0	1623.0	**485.7**	60.1	73.7	75.8	80.6	**50.4**	NA
18 SigLIP2 | Qwen-3-8B	**82.5**	**1648.4**	453.6	**62.2**	**77.2**	**77.7**	**82.6**	49.1	**48.3**
19 SigLIP2 | Lilium-1B	64.7	1383.5	278.9	39.0	57.4	66.4	63.9	35.4	45.6
20 SigLIP2 | Lilium-4B	75.5	1484.8	334.6	47.1	61.8	69.7	74.8	37.8	NA
21 SigLIP2 | Lilium-8B	77.4	1439.2	335.4	51.4	71.4	71.5	76.9	42.3	NA
22 SigLIP | Gemma3-4B	78.3	1617.9	433.2	54.9	69.8	76.6	80.7	43.8	43.5
Open Source									
23 SigLIP | Qwen2-7B *LLaVA-OV*	76.4	1537.4	439.6	55.4	27.9	71.0	80.0	46.4	**50.8**
24 Qwen2.5ViT | Qwen2-7B *Qwen2.5-VL*	81.9	1677.7	654.6	**63.1**	**32.8**	**82.9**	82.8	50.9	40.6
25 Qwen3ViT | Qwen3-8B *Qwen3-VL*	**84.0**	**1742.1**	**660.7**	62.2	26.6	80.9	**84.0**	**52.4**	47.6
26 SigLIP | Gemma3-4B *Gemma3*	67.9	1202.1	398.6	36.5	11.4	62.1	71.2	39.7	34.7

Table 2: **Public multimodal tasks comparison across model architectures and sizes.** We report performance of vision encoder and LLM combinations on public evaluation sets, we also report the split or metric in parenthesis (top row). Internal E-commerce Adaptation models indicate VLMs fully trained top to bottom starting from pre-trained backbones, Open Source indicates models not trained by us, the original *model names* are next to their architectural structure and *NA* indicates results not available. Higher is better (↑).

16), the results are inconclusive, as there is no clear winner between the two encoders. This highlights the complicated relationship with the image modality and task definition, which we will also discuss below for the item intelligence task. For example, the native resolution feature of the *Qwen2.5ViT* might be beneficial for tasks like aspect prediction, where small image details might be important, however we observe weaker results in more reasoning-oriented results in tasks like fashion understanding.

4.4 Importance of Text-Decoder

Comparing the impact of different LLMs when used as backbone with same vision encoder, we observe an influence of (a) domain knowledge of the LLM, (b) general knowledge and (c) model size, which we detail next.

E-commerce Knowledge Helps We compare VLMs based on **Llama-3.1 8B** against the **e-Llama3.1-8B** and **Lilium-8B** variants on the general-domain benchmarks (see Table 2 rows 14, 15, 21), with similar performance. This makes sense, as the underlying text-only LLMs do perform similar on general-domain text-based benchmarks as well. However, when looking at e-commerce specific performance (see Table 1 rows 1, 2, 8) we find that the e-commerce knowledge of e-Llama and Lilium leads to a better adaptability. Table 2, also shows on eComMMMU substantial gains for Gemma3-4B (row 22 vs 26). Due to time constraints we were not able to fully evaluate all models (hence the NA), also we evaluate the use slightly differently metrics than the paper. A deeper discussion in Appendix A.8.

General Capability Helps To see if and how the general-domain capabilities of the text decoder influence final performance, we compare **Qwen3** and **Gemma3** models against previous generation **(e)-Llama** and **Lilium**. The former are trained on significantly more data, therefore they exhibit higher performance on general domain text-only benchmarks. Generally, looking at Table 2, and also comparing model sizes, we find that better capabilities of the text-decoder help improve performance on general domain VLM benchmarks. More interestingly, we find that they also lead to improvements on some e-commerce specific tasks (see Table 1), especially Aspect Prediction. Together with the findings from Section 4.4, this leads us to believe that further gains are possible using a domain-adapted version of the Qwen3/Gemma3 text-decoders, which we leave to future work.

Model Size: Important for Some Tasks Investigating the effect of the size of the text-decoder, we find a consistent trend across both general-domain (Table 2) and e-commerce-specific domain (Table 1). In both cases, larger models lead to stronger performance. However, there seems to be a task-depended threshold for which just increasing model size no longer seems to help. For example, for the Fashion subset of the Aspect Prediction task, going from 1 billion to 4 billion parameters parameters leads to improvements, while going from 4 billion to 8 billion does not. The latter is also consistent for both Lilium and the Qwen3 model families. A

Model Name		Multi-Image Item Intelligence					
		f1-score (↑)	precision (↑)	recall (↑)	verifiable-correct (↑)	verifiable-incorrect (↓)	unverifiable (↓)
0-shot							
27 Gemma3 4B		32.8	33.1	**36.7**	53.6	21.3	25.1
28 Gemma3 27B	*primary image only*	25.5	52.1	18.3	71.6	24.5	3.9
29 Gemma3 27B		**44.8**	**61.8**	36.6	**80.4**	**15.9**	**3.8**
Finetuned							
30 SigLIP2 | e-Llama3.1-8B		42.5	57.0	35.3	72.0	24.0	4.0
31 Qwen2.5ViT | e-Llama3.1-8B		28.7	60.4	20.4	72.2	26.0	1.9
32 Qwen2.5VL-7B		29.3	62.9	20.6	75.3	23.0	1.7
33 Llama-3.1-Nemotron-Nano-VL-8B-V1		50.9	63.3	44.0	79.2	18.9	1.9
34 Gemma3 4B		50.5	64.9	42.8	79.4	17.1	3.5
35 Gemma3 12B		51.8	67.7	43.5	**81.3**	15.7	**3.1**
36 Gemma3 27B		**52.6**	**68.0**	**44.6**	81.2	**15.2**	3.6
Finetuned with Better Labels							
37 Gemma3 4B		53.8	68.1	49.6	82.7	15.9	2.0
38 Gemma3 12B		58.2	71.2	50.9	84.2	14.0	1.7
39 Gemma3 27B		**58.8**	**71.0**	**51.9**	**85.2**	**13.1**	**1.6**
40 Gemma3 4B	*pan&scan*	56.9	68.3	50.5	83.1	15.1	1.8
41 Gemma3 4B	*image crops*	58.0	69.5	51.5	84.7	13.7	**1.6**

Table 3: **Multi-Image Item Intelligence Comparison.** We report performance of different models on multiple types of finetuning strategies (0-shot, Finetuned, Finetuned with Better Labels) over our multi-image item intelligence benchmark. The *italic* next to the model names indicates different inference strategy.

similar trend can be seen on MME.

4.5 Item Intelligence

The Item Intelligence task extracts attributes targeted at regulatory compliance questions. Our baseline is a non-customized Gemma3-27B. In our experiments, we show how we greatly improve quality and efficiency by fine-tuning on this task, while obtaining further improvements by modeling for task-specific characteristics.

Single vs Multi-image We start by establishing the 0-shot performance of the **Gemma3-27B** VLM on the item intelligence task. We compare two settings: (i) the model is given just the primary image of the corresponding listing (ii) the model is given the full set of images. From Table 3 row 28 & 29, we can see that it is definitely beneficial for the model to have access to all existing images of a listing. We also test the performance of the more efficient **Gemma3-4B** model (row 27), but find the model predictions to be of worse quality.

Fine-Tuning Helps Next, we compare fine-tuning a model and compare against the zero-shot approach from Section 4.5. We fine-tune a subset of the models we discussed above for the general e-commerce adaptation. As can be seen in Table 3 row 36, fine-tuning significantly improves performance of the **Gemma3-27B** model. Furthermore, performance of the much smaller **Gemma3-4B** VLM (row 34) is also strong after fine-tuning. Other models like **Qwen2.5ViT | e-Llama3.1-8B** and **Qwen2.5VL-7B (ft)** fall behind. Another big advantage of fine-tuning is the greatly improved inference efficiency. Due to smaller model size and shorter prompt size, we achieve ca. 3.8x inference speedup when replacing Gemma3-27B with the finetuned Gemma3-4B model (see Table 4).

It Matters Where You Look In an effort to further improve results, we test the image bounding boxes approach outlined in Section 3.2.3, which leads to better labels for training examples. As can be seen from Table 3 rows 37 - 39, this approach leads to significant improvements for all model sizes. We also test including the image crops in inference (row 41) and compare against the 'Pan & Scan' feature from Gemma3 (row 40). We find that both approaches improve performance, but our more targeted cropping leads to stronger results.

5 Conclusion

We introduced a reproducible, backbone-agnostic recipe for adapting open-weight VLMs to the attribute-centric, multi-image, and noisy characteristics of e-commerce. To evaluate this, we constructed a benchmark suite spanning Aspect Prediction, Deep Fashion Understanding, Dynamic Attribute Extraction and multi-image Item Intelligence. Across extensive ablations we show how targeted adaptation can deliver substantial in-domain gains while preserving broad capabilities. Lastly, in a production-style Item Intelligence case study, targeted cropping plus improved labels and fine-tuning yielded strong quality gains and multiple times faster inference compared to general-purpose VLMs.

6 Limitations

Our study has the following limitations.

- **(i) Monolingual scope.** All model adaptation, supervision, and evaluation were conducted in English. Consequently, we do not characterize cross-lingual transfer to product ontologies, attribute surface forms, or unit/size conventions that are language– and locale-specific (i.e., multi-script OCR for size charts, EU/JP sizing, or currency/decimal formats).

- **(ii) Platform dependence.** The instruction corpus and benchmarks are sourced predominantly from a single marketplace, and many prompts/targets were curated or verified via automated pipelines. This creates potential distributional coupling to that platform's taxonomy, seller conventions, imaging styles (studio vs. user-generated), and metadata density. This hinders portability to other marketplaces with different attribute schema or listing norms remains uncertain.

- **(iii) LLM-mediated supervision and evaluation.** Portions of training signals (i.e., pseudo-labels, instruction filtering) and some evaluations rely on LLMs. This introduces annotator bias, style bias, and measurement noise; moreover, evaluator–model family overlap can inflate or deflate measured gains due to inductive-bias alignment in "LLM-as-judge" scenarios.

- **(iv) Coverage of phenomena.** While broad, our evaluation is not exhaustive: the Dynamic Attribute Extraction (DAE) set is $\sim$ 1k examples and category coverage emphasizes selected fashion and high-volume verticals. As a result, performance on long-tail categories, rare attributes, region-specific variants, heavily composited images, or atypical listing styles is under-constrained. Overall, the reported improvements should be interpreted as evidence of promise under these conditions rather than as guarantees of cross-lingual or cross-platform robustness.

- **(v) Long Image Sequence Handling.** In scenarios with more than 10 images (rare), we noticed our models may suffer from Out-of-Memory (OOM) issues as well as long inference times. This is particularly tricky for Multi-Image Item Intelligence and eComMMMU benchmarks. While having 10 or more images is rare, this can lead to issues in potential production use-cases. While this could be solved by training larger context LLMs or through token efficient strategies (Zhang et al., 2025), it is something worth addressing in the future.

References

2024. Gpt-4o system card.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Haibin Chen, Kangtao Lv, Chengwei Hu, Yanshi Li, Yujin Yuan, Yancheng He, Xingyao Zhang, Langming Liu, Shilei Liu, Wenbo Su, and Bo Zheng. 2025a. Chineseecomqa: A scalable e-commerce concept evaluation benchmark for large language models. *Preprint*, arXiv:2502.20196.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, abs/2311.16079.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 32 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *Preprint*, arXiv:2409.17146.

Wenxuan Ding, Weiqi Wang, Sze Heng Douglas Kwok, Minghao Liu, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Junxian He, and Yangqiu Song. 2024. Intentionqa: A benchmark for evaluating purchase intention comprehension abilities of language models in e-commerce. *Preprint*, arXiv:2406.10173.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Gemma-Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. 2019. Amazonqa: A review-based question answering task. *Preprint*, arXiv:1908.04364.

Christian Herold, Michael Kozielski, Tala Bazazo, Pavel Petrushkov, Seyyed Hadi Hashemi, Patrycja Cieplicka, Dominika Basaj, and Shahram Khadivi. 2025. Domain adaptation of foundation llms for e-commerce. *Preprint*, arXiv:2501.09706.

Christian Herold, Michael Kozielski, Leonid Ekimov, Pavel Petrushkov, Pierre-Yves Vandenbussche, and Shahram Khadivi. 2024. Lilium: ebay's large language models for e-commerce. *Preprint*, arXiv:2406.12023.

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, Haodong Wang, Zhengyang Wang, Wenju Xu, Jingfeng Yang, Qingyu Yin, Xian Li, Priyanka Nigam, Yi Xu, Kai Chen, and 3 others.

2024. Shopping mmlu: A massive multi-task online shopping benchmark for large language models. *Preprint*, arXiv:2410.20745.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, and 1 others. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

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

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, and 1 others. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Yunxin Li, Baotian Hu, Wenhan Luo, Lin Ma, Yuxin Ding, and Min Zhang. 2024b. A multimodal in-context tuning approach for e-commerce product description generation. *Preprint*, arXiv:2402.13587.

Xinyi Ling, Hanwen Du, Zhihui Zhu, and Xia Ning. 2025. Ecommmmu: Strategic utilization of visuals for robust multimodal e-commerce models. *Preprint*, arXiv:2508.15721.

Xinyi Ling, Bo Peng, Hanwen Du, Zhihui Zhu, and Xia Ning. 2024. Captions speak louder than images (caslie): Generalizing foundation models for e-commerce from high-quality multimodal instruction data. *Preprint*, arXiv:2410.17337.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Langming Liu, Haibin Chen, Yuhao Wang, Yujin Yuan, Shilei Liu, Wenbo Su, Xiangyu Zhao, and Bo Zheng. 2025. Eckgbench: Benchmarking large language models in e-commerce leveraging knowledge graph. *Preprint*, arXiv:2503.15990.

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. 2024b. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, and 13 others. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *Preprint*, arXiv:2403.09611.

Mistral AI. 2024. Mistral small 3: Mistral's most efficient 24b model. Accessed: 2024-10-31.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenGVLab-Team. 2024. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy.

Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data. *Preprint*, arXiv:2402.08831.

QwenTeam. 2025. Qwen3-vl: Sharper vision, deeper thought, broader action.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping queries dataset: A large-scale ESCI benchmark for improving product search.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, and 6 others. 2023. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, and 7 others. 2024. Climategpt: Towards AI synthesizing interdisciplinary research on climate change. *CoRR*, abs/2401.09646.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Preprint*, arXiv:2406.16860.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. 2024. Locca: Visual pretraining with location-aware captioners. *Preprint*, arXiv:2403.19596.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

Wei Xue, Zongyi Guo, Baoliang Cui, Zheng Xing, Xiaoyi Zeng, Xiufei Wang, Shuhui Wu, and Weiming Lu. 2024. Pumgpt: A large vision-language model for product understanding. *Preprint*, arXiv:2308.09568.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Christoph Zauner. 2010. Implementation and benchmarking of perceptual image hash functions. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg Campus, Hagenberg, Austria.

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, and 4 others. 2024. Mm1.5: Methods, analysis & insights from multimodal llm fine-tuning. *Preprint*, arXiv:2409.20566.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. *Preprint*, arXiv:2501.03895.

A Appendix

A.1 Related Work (Continued)

Multi Purpose MLLMs Since the advent of Visual Instruction Tuning (Liu et al., 2023), many have grasped the impact of combining CLIP Vision Encoders (Radford et al., 2021) with Large Language Models (LLMs) (Radford et al., 2019; Chiang et al., 2023; Touvron et al., 2023; Dubey et al., 2024) to enable cross modality understanding with LLMs. Most notably LLaVA (Liu et al., 2023) and GPT4V (OpenAI et al., 2024), have paved the way for more diverse and varied MLLMs. Recent investigations have advanced along several complementary fronts. From a systematical decomposition of the training pipeline and characterization of model behavior across a variety of pre-trained backbones (McKinzie et al., 2024; Zhang et al., 2024; Laurençon et al., 2024). to the efficient processing of images spanning multiple resolutions (Liu et al., 2024a; Wang et al., 2024; OpenGVLab-Team, 2024) as well as the development of fully open multimodal foundation models (Deitke et al., 2024). Multimodal Large Language Models have consistently achieved state-of-the-art results across a broad spectrum of downstream applications, encompassing image captioning (Yu et al., 2022; Chen et al., 2023a; Wan et al., 2024), visual question answering (Liu et al., 2024a), image understanding (Liu et al., 2023; Tong et al., 2024), and complex reasoning tasks (Xu et al., 2024).

E-commerce Model Adaptation General-domain pretrained LLMs often struggle with domain-specific tasks, motivating domain-specific pretraining or targeted domain adaptation (Lewkowycz et al., 2022; Chen et al., 2023b; Rozière et al., 2023).

Pretraining a domain-specific LLM from scratch results in the highest degree of adaptation, including domain-specific knowledge, vocabulary, and more (Wu et al., 2023; Li et al., 2023; Herold et al., 2024). However, it is also extremely costly and slow, and requires a huge amount of domain-specific data.

As an alternative, continuous pretraining on in-domain text or fine-tuning an existing model can also substantially boost performance on domain-specific tasks (Azerbayev et al., 2024; Shao et al., 2024; Thulke et al., 2024; Herold et al., 2025), at the cost of less overall customizability.

Figure 3: **Visual Breakdown of our benchmarks.** We choose four representative examples from each of our proposed benchmarks to showcase the tasks.

A.2 General Domain Multimodal Benchmarks

To evaluate our models on existing e-Commerce tasks we choose eComMMMU (Ling et al., 2025), one of the few comparing evaluation suits for MLLMs in online shopping. It is comprised of over 35k multi-image samples spanning over 8 tasks. Furthermore, we employ 8 other general multimodal understanding benchmarks, ensuring close monitoring of general performance. These are MM-Bench (Liu et al., 2024b) covering object detection, text recognition, action recognition, among many others, MMMU (Yue et al., 2024) evaluating Mulimodal LLMs on perception, knowledge, and reasoning, CVBench (Tong et al., 2024) evaluating visual-centered capabilities of our models, and finally, MME (Fu et al., 2024), a comprehensive benchmark dividing between perception and cognition tasks, with 15 subcategories. AI2D (Kembhavi et al., 2016) a Diagram/ChartQA with 3,009 examples, and MMStar (Chen et al., 2024) 1.5k samples across 6 categories (Perception, Math, Science & Tech, Logical, Instance Reasoning). TextVQA (Singh et al., 2019) designed to stress-test capabilities of VQA models in OCR, with 5k examples.

A.3 Methodology

A.4 Our E-commerce Benchmarks

Aspect Prediction We propose our Aspect Prediction evaluation suite. This set is divided into three different sub-parts, each tasked with a specific objective. The first set is comprised of 2600 general aspect prediction questions on almost all e-commerce categories (collectibles, car parts, cards, fashion, etc...). In the last two, we evaluate the model's ability to predict aspects in Fashion, setting with and without additional textual contexts provided by item title and category, both with 1600 examples. All three are evaluated through string matching after post-processing. Although online shopping is often dominated by fashion items, we deem important to include evaluation sets which could more accurately capture the broad spectrum of online marketplaces.

Multi-image item intelligence Many attributes related to product safety and compliance such as certifications, ingredients, warning labels are not provided by the item's seller, and manual inspection is inherently slow and costly. To address this, we propose a structured set designed to systematically extract and normalize visible information into consistent JSON outputs, enabling streamlined verification and recall matching processes. Our benchmark prioritizes product categories with prominent packaging and labeling signals, including toys, electronics, appliances, cosmetics, supplements, batteries, PPE, and food items. It handles diverse image sources such as product listing galleries, detailed zoomed-in views, and user-uploaded photographs. The resulting structured schema encompasses essential data elements such as *Product Identifiers*, *Product Attributes*, *Product Origin*, and *Regulatory Safety*, ensuring accurate and consistent outputs. We evaluate through LLM-as-a-judge.

Deep Fashion Understanding Characterizing complex fashion features is a fundamental component of e-commerce assistants. To accurately evaluate deep fashion understanding, we designed a specialized sub-benchmark consisting of 3k samples divided into four distinct subsets: *Apparel Men Shirts*, *Apparel Women Tops*, *Handbags*, and *Sneakers*. Each subset targets critical attributes relevant to the product type, structured into clear classification categories. For instance, Apparel Men Shirts are evaluated based on Sleeve Length, Neckline, Pattern, and Color, with predefined classes such as 'Short Sleeve', 'Crew Neck', 'Striped', and 'Orange'. Apparel Women Tops share similar but more extensive attribute categories, including additional neckline and pattern options like 'Off the Shoulder' and 'Paisley'. Handbags and Sneakers subsets specifically focus on accurately identifying brand labels, such as 'Louis Vuitton' or 'Nike'. Evaluation involves prompting the model to categorize items precisely according to the provided attribute classes.

Dynamic Attribute Extraction Extracting visual item attributes from an image is a complicated yet essential task. This evaluation set benchmarks a model's ability to enumerate and structure all visually grounded attributes from an image without a predefined schema. Each instance is prompted only once, requiring the model to decide which properties are salient, choose attribute names, and serialize values as key–value pairs (e.g., format, edition, material, artist, counts, genres, brand, model). The benchmark comprises 1,000 synthetically generated with GPT-4o (gpt, 2024), human-verified examples and emphasizes attributes that are strictly supported by the pixels. Unlike fixed-ontology extraction, Dynamic Attribute Extraction (DAE) stresses e-Commerce generalization by incentivizing exhaustive yet faithful outputs, avoiding hallucinated fields. A typical response for a text-rich object, such as a DVD cover, would be a compact JSON record as show in Appendix 3. By design, DAE probes the practical skill needed in cataloging, document understanding, and product intelligence workflows where schemas are fluid and attributes must be discovered on the fly.

A.5 Item Intelligence Fine-tuning

Using both the original images and all derived crops for inference is computationally expensive, as the Gemma-3 image encoder assigns a fixed 256 visual tokens per image, causing inference cost to scale linearly with the number of images, even when many of them are small. On our training dataset, this resulted in a median of 12 and a maximum of 43 images per item. To address this, we construct crops covering the regions of interest optimized for the Gemma-3 encoder by identifying the smallest enclosing square that covers all bounding boxes, consistent with the model's square image format. Finally, we apply a lightweight deduplication step using perceptual hashing (pHash) (Zauner, 2010), reducing the number of images per item to a median of four and a maximum of nine.

A.6 Inference Speed Comparison

Model	sec/example
0-shot	
Gemma 27B	25.5
Finetuned	
Gemma 27B	19.3
Gemma 4B	6.7

Table 4: **Inference speed comparison.** We report the speeed comparison on the Multi-Image Item Intelligence task between the 0-shot Gemma 27B model and the 4B and 27B finetuned variants. Experiments were conducted on a single A100 GPU using a recent version of vLLM.

A.7 Our Approach to E-commerce Adaptation

```
Our mid-stage datasets:
  - json_path: ./llava_ov/LLaVA-ReCap-558K.
    ↪ json
    sampling_strategy: all
  - json_path: ./llava_ov/LLaVA-ReCap-118K.
    ↪ json
    sampling_strategy: all
  - json_path: ./llava_ov/LLaVA-ReCap-CC3M.
    ↪ json
    sampling_strategy: all
  - json_path: ./llava_ov/
    ↪ synthdog_en_processed.json
    sampling_strategy: all
```

```
Our single-image LLaVA-OneVision sets for
    ↪ visual instruction tuning:
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_mavis_math_metagen.json
    sampling_strategy: "first:10%"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_mavis_math_rule_geo.json
    sampling_strategy: "first:10%"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_VisualWebInstruct(
    ↪ filtered).json
    sampling_strategy: "all"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_chrome_writting.json
    sampling_strategy: "first:20%"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_iiit5k.json
    sampling_strategy: "first:20%"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_hme100k.json
    sampling_strategy: "first:10%"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_orand_car_a.json
    sampling_strategy: "first:10%"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_llavar_gpt4_20k.json
    sampling_strategy: "first:10%"
  - json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_ai2d(gpt4v).json
    sampling_strategy: "all"
  - json_path: ./llava_ov/meta_ov/LLaVA-
```

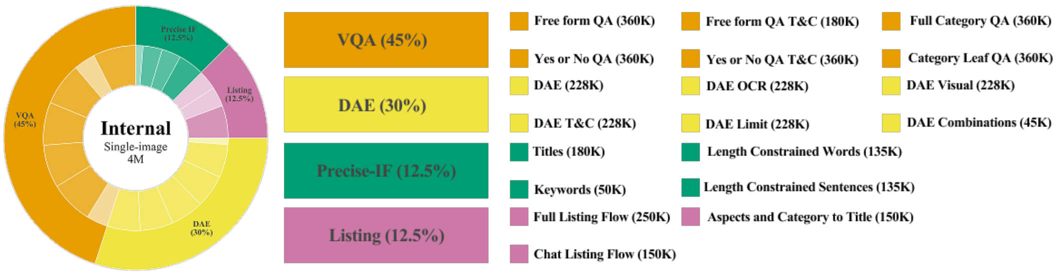

Figure 4: **Internal Single-Image Visual Instruction Tuning Set.** We break down the components of our internal single-image instruction tuning set. The pie chart on the left shows the percentages of tasks in our set. On the right we breakdown each tasks with its own sub tasks with the total number of instructions in parenthesis.

```
   ↪ OneVision-Data_infographic_vqa.json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_infographic(gpt4v).json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_lrv_chart.json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_lrv_normal(filtered).
   ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_scienceqa(nona_context).
   ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_allava_instruct_vflan4v.
   ↪ json
  sampling_strategy: "first:30%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_allava_instruct_laion4v.
   ↪ json
  sampling_strategy: "first:30%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_textocr(gpt4v).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_ai2d(internvl).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_textcaps.json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_ureader_cap.json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_ureader_ie.json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_vision_flan(filtered).
   ↪ json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_mathqa.json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_geo3k.json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_geo170k(qa).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_geo170k(align).json
```

```
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_sharegpt4o.json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_sharegpt4v(coco).json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_sharegpt4v(knowledge).
   ↪ json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_sharegpt4v(llava).json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_sharegpt4v(sam).json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_CLEVR-Math(MathV360K).
   ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_FigureQA(MathV360K).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_Geometry3K(MathV360K).
   ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_GeoQA+(MathV360K).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_GEOS(MathV360K).json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_IconQA(MathV360K).json
  sampling_strategy: "first:5%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_MapQA(MathV360K).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_PMC-VQA(MathV360K).json
  sampling_strategy: "first:1%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_Super-CLEVR(MathV360K).
   ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_TabMWP(MathV360K).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
   ↪ OneVision-Data_UniGeo(MathV360K).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
```

```
    ↪ OneVision-Data_VizWiz(MathV360K).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_image_textualization(
    ↪ filtered).json
  sampling_strategy: "first:20%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_ai2d(cauldron,
    ↪ llava_format).json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_chart2text(cauldron).
    ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_chartqa(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_diagram_image_to_text(
    ↪ cauldron).json
  sampling_strategy: "all"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_hateful_memes(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_hitab(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_iam(cauldron).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-
    ↪ Data_infographic_vqa_llava_format.json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_intergps(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_mapqa(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_rendered_text(cauldron).
    ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_robut_sqa(cauldron).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_robut_wikisql(cauldron).
    ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_screen2words(cauldron).
    ↪ json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_tabmwp(cauldron).json
  sampling_strategy: "first:5%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_tallyqa(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:5%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_st_vqa(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_visual7w(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_visualmrc(cauldron).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_vqarad(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_vsr(cauldron,
    ↪ llava_format).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_vistext(cauldron).json
  sampling_strategy: "first:10%"
- json_path: ./llava_ov/meta_ov/LLaVA-
    ↪ OneVision-Data_websight(cauldron).json
  sampling_strategy: "first:10%"
```

A.8 Experiments

eComMMMU Given the similar goals of eComMMMU (Ling et al., 2025) and our work, we decided to include it within our general benchmarks. As it is apparent from Table 2, there are some inconsistencies in our evaluations. First, due to time constraints we were not able to fully evaluate all models on this benchmark. This is why 1 and 5 show only a subset of models. Secondly, we made amendments on (a.) the amount of images for each example and (b.) the final Average metric. For (a.) the eComMMMU paper uses either the main image or an (automatically) relevance-filtered subset which is not public. We first tried to include all images but hit Out-of-Memory issues. Some test-set examples contained north of 10 images. Due to our models context-sizes, we could not concurrently consider samples with more than 10 images. Thus we capped the amount of images to 10 removing all excess, but keeping all textual examples. The second (b.) was a design choice on our side. We wanted to avoid to use the 'average model rank' for reproducibility and reporting purposes. We thus performed a weighted average across all tasks. This is what is shown in Table 1 and as Avg. in Table 5.

Our hope is that we can overcome the scale issues for the final version of the paper, and contribute to an evaluation setting that is as conducive as possible to third-party model evaluations usind eComMMMU as a benchmark for multimodal LLMs.

Vision Encoder | LLM		eComMMMU (GTS)								
		AP	BQA	CP	SR	MPC	PSI	SA	PRP	Avg.
Internal E-commerce Adaptation										
42 SigLIP2 | In-house-LLM-A		36.7	19.5	50.1	4.2	63.3	43.3	64.3	42.5	46.1
43 Qwen2.5ViT | In-house-LLM-A		52.5	18.8	49.5	4.6	61.6	65.2	76.6	39.7	45.9
44 SigLIP2 | Qwen-3-8B		61.8	35.6	50.7	8.4	64.9	31.6	70.2	20.4	48.3
45 SigLIP2 | In-house-LLM-B-1B		33.5	17.7	50.5	4.5	51.7	77.1	12.5	51.9	45.6
46 SigLIP | Gemma3-4B		59.3	34.8	51.1	6.7	63.9	26.0	50.2	13.8	43.5
Open Source										
47 SigLIP | Qwen2-7B	*LLaVA-OV*	33.7	20.5	50.5	5.6	65.1	76.8	34.7	50.3	50.8
48 Qwen2.5ViT | Qwen2-7B	*Qwen2.5-VL*	31.2	46.2	32.5	10.0	65.7	26.9	58.0	37.0	40.6
49 Qwen3ViT | Qwen3-8B	*Qwen3-VL*	54.3	38.6	52.4	11.9	64.2	30.4	73.0	26.5	47.6
50 SigLIP | Gemma3-4B	*Gemma3*	45.2	32.5	50.3	11.0	39.7	29.9	49.0	14.6	34.7

Table 5: **eComMMMU Full sub-tasks results.** We report performance of different models on eComMMMU benchmark the GTS subset with *multiple* image per sample. We show performance on all sub-tasks (AP = answerability prediction , BQA = binary question answering , CP = click through prediction, SR = sequential recommendation, MPC = multiclass product classification, PSI = production substitute identification, PRP = product relation prediction, SA = sentiment analysis). For SR we report the Recall@1 score, whereas for all others accuracy. The Average (Avg) is calculated weighting based on the amount of samples per sub-task taking SR into account as well. The *italic* next to the model names indicates different inference strategy.