

Captions Speak Louder than Images: Generalizing Foundation Models for E-commerce from High-quality Multimodal Instruction Data

Xinyi Ling¹, Hanwen Du¹, Bo Peng¹, Zhihui Zhu¹, Xia Ning^{1 2 3}✉

¹Department of Computer Science and Engineering, The Ohio State University

²Translational Data Analytics Institute, The Ohio State University

³Department of Biomedical Informatics, The Ohio State University

{ling.303, du.1128, peng.707, zhu.3440, ning.104}@osu.edu

Abstract

Multimodal foundation models (MFMs) have demonstrated strong capabilities in e-commerce by effectively leveraging multimodal data to enhance product understanding and user experience. However, the development of e-commerce MFMs is hindered by two challenges: (1) the scarcity of large-scale, high-quality multimodal benchmark datasets; and (2) the lack of effective multimodal information integration methods in e-commerce. To address these challenges, we introduce MMECInstruct, the first large-scale, high-quality multimodal instruction dataset designed specifically for e-commerce MFMs. MMECInstruct comprises 75,000 samples covering 7 real-world e-commerce tasks, supporting both in-domain (IND) and out-of-domain (OOD) evaluations. Leveraging MMECInstruct, we develop CASLIE, a lightweight framework that enhances multimodal information understanding and integration for e-commerce. Our comprehensive evaluation demonstrates that MMECInstruct endows CASLIE with advanced capability and strong generalizability in e-commerce applications. MMECInstruct and CASLIE models are publicly accessible through <https://github.com/ninglab/CASLIE>.

1 Introduction

Multimodal data, encompassing diverse modes and types of information such as text and images, is ubiquitous and essential for many real-world applications (Antol et al., 2015; Wang et al., 2023; Mu et al., 2024; Chen et al., 2021). In e-commerce, multimodal data is especially important: product content typically combines visual and textual information, and user interactions involve diverse data types across multiple modalities. Effectively harnessing multimodal data for e-commerce exhibits strong promise to allow for a more comprehensive depiction of product attributes and uncover deeper

insights into customer preferences, which single-modal data alone may not suffice (Wang et al., 2023; Peng et al., 2023). With the recent surge of Large-Language Models (LLMs) on e-commerce tasks and their remarkable performance (Peng et al., 2024; Li et al., 2024b; Shi et al., 2023), multimodal data are expected to drive new breakthroughs in e-commerce applications, together with the development of Multimodal Foundation Models (MFMs).

However, despite the richness of multimodal e-commerce data, there are significant challenges that hinder its optimal use by foundation models (Wang et al., 2023; Liu et al., 2023c): **(1) Scarcity of large-scale, high-quality multimodal benchmark datasets for a large variety of e-commerce applications.** It is highly nontrivial to curate such a dataset due to the complexity of the data processing involved (e.g., selecting products that possess rich, high-quality data across all modalities). While initiatives for unimodal e-commerce benchmark datasets for LLMs have been undertaken (Peng et al., 2024; Li et al., 2024b; Shi et al., 2023), to the best of our knowledge, no such multimodal counterparts exist. **(2) Lack of effective multimodal information integration methods for e-commerce tasks.** Current LLM-based e-commerce models (Peng et al., 2024; Li et al., 2024b) often focus predominantly on one modality, typically text. Existing multimodal approaches (Chia et al., 2022; Yu et al., 2022) attempt to map different modalities into a shared latent space, following the CLIP paradigm (Radford et al., 2021) developed from the computer vision domain. However, this alignment-based strategy overlooks key challenges unique to e-commerce.

First, multimodal information often complements rather than aligns (Lin et al., 2025; Dufumier et al., 2025; Baldrati et al., 2022), while alignment is a core assumption in CLIP. For instance, an image of a large shampoo bottle conveys information about its bottle size but not its fragrance, while user

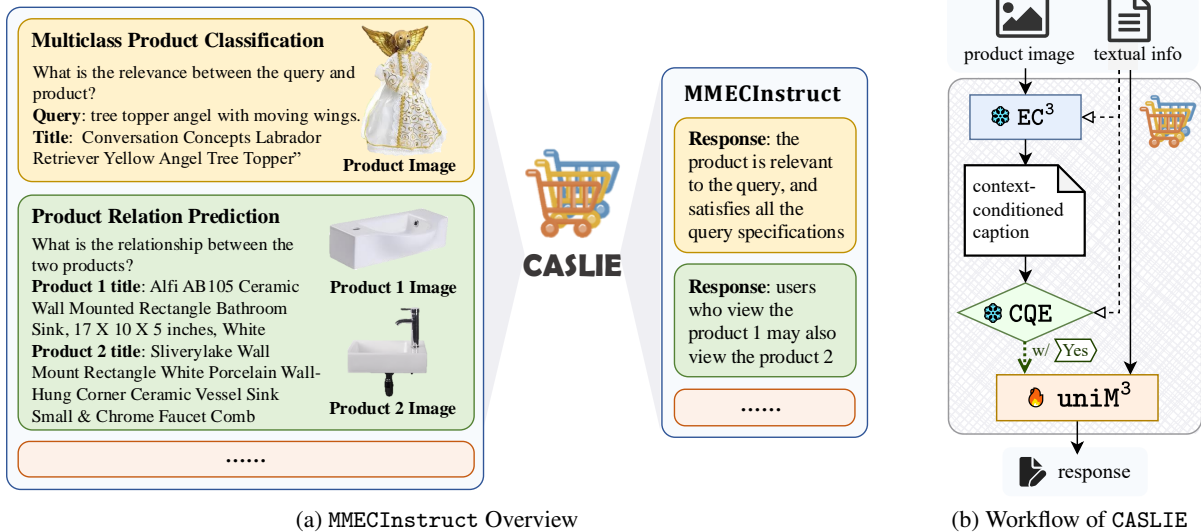


Figure 1: MMECInstruct and CASLIE overview

reviews may praise its fragrance. Thus, image and user reviews are complementary to each other. Second, the relevance of visual information is highly context-dependent: the same image feature may be crucial in one product category but irrelevant in another (Li et al., 2014; Gu and Deng, 2024).

To address these challenges, we introduce **MMECInstruct**, the first-ever, large-scale, and high-quality multimodal instruction dataset designed specifically for e-commerce applications. As shown in Figure 1a, MMECInstruct consists of 75,000 samples spanning 7 widely-performed real-world e-commerce tasks. Each data sample includes an instruction, one or multiple images, a textual input, and an expected response, enabling the development and evaluation of e-commerce foundation models. MMECInstruct is carefully curated to support a broad range of experimental settings, including in-domain (IND) evaluation for all 7 tasks, out-of-domain (OOD) evaluation (i.e., evaluation task on products of new category not included in the training set) for 5 tasks, and task-specific studies, ensuring robustness in real-world scenarios. We perform rigorous processing to ensure the high quality of the MMECInstruct.

Leveraging MMECInstruct, we develop **CASLIE** – **C**Aptions **S**peak **L**ouder than **I**mage**S**, a simple, lightweight, yet effective learning framework for e-commerce MFMs, which integrates text and images for e-commerce tasks. Figure 1b shows the workflow of CASLIE. CASLIE comprises three modules: (1) a context-conditioned caption generation module, denoted as EC^3 , that translates images into captions conditioned on given context,

(2) a caption quality evaluation module, denoted as CQE, that excludes ineffectual visual information, and (3) a modality information fusion module, denoted as $uniM^3$, that seamlessly integrates visual and textual information for downstream tasks. CASLIE processes images in a way that adapts to product-specific contexts, generating high-quality captions that bridge visual and text in a context-aware way, making it fundamentally different from previous work (Chia et al., 2022; Liu et al., 2023b).

Existing MFMs typically embed and align visual and textual inputs using context-agnostic fusion techniques (Li et al., 2024a). However, they often fail to distinguish helpful content from noise in images, resulting in suboptimal multimodal representations for e-commerce applications. Different from these models, CASLIE offers a simple, light-weight, training-free yet effective fusion framework, enabling a unified view of multimodal data for e-commerce tasks. Another advantage of CASLIE is its plug-and-play design: all its modules can be easily reimplemented when newer and more advanced models become available, allowing for seamless integration of the most suitable options. Our experiments show that CASLIE is significantly empowered with MMECInstruct to outperform state-of-the-art baselines across multiple e-commerce tasks. We make MMECInstruct publicly available at <https://github.com/ninglab/CASLIE> to facilitate further research in multimodal learning for e-commerce.

Mod.	Dataset	Size	Div.	Ins.
Text	Amazon-M2 (Jin et al., 2024a)	3.6M	✗	✗
	Shopping Queries (Reddy et al., 2022)	130K	○	✗
	EcomInstruct (Li et al., 2024b)	2.6M	✓	✓
	ECInstruct (Peng et al., 2024)	116K	✓	✓
	Shopping MMLU (Jin et al., 2024b)	11K	✓	✗
	AmazonQA (Gupta et al., 2019)	924K	○	✗
Text & Image	MEP-3M* (Liu et al., 2023a)	3M	✗	✗
	MMECInstruct (ours)	75K	✓	✓

Table 1: Comparison with existing e-commerce dataset. “Mod.” denotes the type of modalit(ies) in the dataset. “Size” denotes the number of samples in each dataset. “Div.” denotes whether the dataset contains diverse tasks. “Ins.” denotes whether the dataset contains instructions for LLM finetuning. *MEP-3M is composed of product meta information, lacking structured formulation for downstream applications. ○The datasets only contain query/QA-related tasks.

2 Related Work

E-commerce Benchmark Developing MFMs for e-commerce requires high-quality datasets that integrate multimodal information. Several existing datasets focus on text-based e-commerce tasks, such as EcomInstruct (Li et al., 2024b) and ECInstruct (Peng et al., 2024), which provide instruction-based learning resources but lack image data, limiting their applicability for multimodal learning. Other datasets, such as Amazon-M2 (Jin et al., 2024a) and the Shopping Query Dataset (Reddy et al., 2022), contain large-scale e-commerce interactions but primarily focus on user behavior and query-related tasks without multimodal coverage. While MEP-3M (Liu et al., 2023a) incorporates both text and image modalities, it lacks structured instructions, making it less suitable for fine-tuning instruction-following multimodal models. In contrast, MMECInstruct *is the first multimodal instruction dataset for e-commerce, offering task-specific, high-quality image-text pairs across seven diverse e-commerce applications.* By addressing these gaps, MMECInstruct establishes a new benchmark for multimodal e-commerce research, enabling robust evaluation and generalization of foundation models.

Multimodal Learning for E-commerce In recent years, remarkable advancements in multimodal learning (Radford et al., 2021; Li et al., 2021; Alayrac et al., 2022; Stevens et al., 2024) have enabled significant progress in integrating vision and language into e-commerce models. For example, CommerceMM (Yu et al., 2022) learns mul-

timodal representations for various e-commerce tasks by aligning paired data from different modalities via contrastive learning. ECLIP (Jin et al., 2023) and FashionCLIP (Chia et al., 2022) introduce CLIP (Radford et al., 2021)-based contrastive pre-training frameworks to learn multimodal e-commerce data representations transferable to downstream tasks. However, CLIP-based models generate image representations from the entire image in a context-free manner, making it difficult to emphasize specific image details conditioned on the given context. In contrast, CASLIE *generates context-conditioned textual representations for images (e.g., captions), highlighting different details depending on the context.* Additionally, CASLIE leverages the world knowledge in MFMs to generate captions, enriching captions with additional information pertinent to target tasks.

3 MMECInstruct Dataset

To advance multimodal learning in e-commerce, we introduce MMECInstruct, a multimodal instruction dataset designed to adapt general-purpose MFMs for e-commerce. MMECInstruct is constructed under three principles: **(1) Multimodality:** Unlike text-only datasets (e.g., EcomInstruct (Li et al., 2024b) and Shopping MMLU (Jin et al., 2024b)), MMECInstruct contains both visual and textual content for each product in various e-commerce tasks, enabling comprehensive multimodal learning of foundation models. **(2) Broad coverage:** MMECInstruct comprises seven diverse and realistic tasks to enable versatile e-commerce modeling and benchmarking (Peng et al., 2024; Jin et al., 2024b,a). **(3) High quality:** The dataset is carefully curated through rigorous validation processes to ensure both accuracy and reliability. As demonstrated in the literature (Hoffmann et al., 2022; Gadre et al., 2024), high-quality instruction-tuning data plays a pivotal role in building powerful foundation models. Figure 1a presents the overview of MMECInstruct, and Table 1 summarizes related e-commerce datasets. More information about MMECInstruct dataset can be found in Appendix A. *To the best of our knowledge, MMECInstruct is the first of its kind.*

3.1 E-commerce Tasks

In line with prior works (Yue et al., 2023; Fang et al., 2024; Peng et al., 2024), MMECInstruct comprises 7 widely-performed real-world e-commerce tasks with real-world data extracted

Task	Definition	Type	Primary Metrics	Data Source
AP	Predict if the product-related question is answerable based on the product information.	Binary classification	F1 score	AmazonQA (Gupta et al., 2019)
CC	Retrieve the category of the product based on the product information.	Retrieval	Recall@1	MAVE (Yang et al., 2022), Amazon Review 2023 (Hou et al., 2024)
PRP	Identify the relationship between two product from “also buy”, “also view”, and “similar”.	Multi-class classification	Macro F1 score	Amazon Review 2023 (Hou et al., 2024)
PSI	Predict if the product can serve as a functional substitute for the user’s query.	Binary classification	F1 score	Shopping Query Dataset (Reddy et al., 2022)
MPC	Given a query and product information, predict relevance between the query and product.	Multi-class classification	Accuracy	Shopping Query Dataset (Reddy et al., 2022)
SA	Identify the sentiment user expressed based on the product review text and review image.	Multi-class classification	Macro F1 score	Amazon Review 2023 (Hou et al., 2024)
SR	Predict the next product that user would be interested in based on user’s purchase history.	Retrieval	Recall@1	Amazon Review 2023 (Hou et al., 2024)

Table 2: Tasks in MMECInstruct dataset

from e-commerce platforms: (1) answerability prediction (AP) (Gupta et al., 2019), (2) category classification (CC) (Yang et al., 2022; Chen et al., 2021), (3) product relation prediction (PRP) (Ni et al., 2019; Xu et al., 2020a), (4) product substitute identification (PSI) (Reddy et al., 2022), (5) multi-class product classification (MPC) (Reddy et al., 2022), (6) sentiment analysis (SA) (Wankhade et al., 2022; Daza et al., 2024), and (7) sequential recommendation (SR) (Li et al., 2023a; Hou et al., 2024; Petrov and Macdonald, 2023). These tasks are designed to cover key functions in modern e-commerce platforms, including search, recommendation, QA, and sentiment analysis. Detailed information about all the e-commerce tasks is presented in Table 2.

3.2 Vision-language Data

Different from existing datasets with text-only instructions (Peng et al., 2024), MMECInstruct includes both visual and textual content for each item. Particularly, the dataset includes (1) product images and user review images as visual information, (2) product titles, product categories, product brands, user queries, user reviews, and user questions as textual content, (3) human-designed structured instructions tailored to real-world scenarios for each task, and (4) ground-truth response to each sample. The multimodal e-commerce data is enriched with synergistic visual and textual inputs, providing a basis for developing and evaluating models on a range of multimodal e-commerce tasks.

3.3 Quality Control

In constructing MMECInstruct, we adopt established principles from other instruction

datasets (Peng et al., 2024; Fang et al., 2024; Yue et al., 2023), focusing on clear instructions, consistent data formatting, and good alignment between input and target outputs (Gadre et al., 2023). Those are critical for training generalizable instruction-following models.

Besides, we exclude products without an accompanying image available to ensure all modalities are consistently available. We select medium-size images (500×500 resolution) for each product to balance visual clarity and computational efficiency. We retain only products that include both detailed textual descriptions and corresponding images to ensure sufficient multimodal information for effective foundation model training. In addition, we remove samples from the test sets that also appear in the training set to prevent data leakage and ensure a clean separation for both IND and OOD evaluations. We further conduct manual scrutiny on the 1,000 randomly sampled instances (Hedt-Gauthier et al., 2013) to ensure the overall data quality of accuracy, clarity, and relevance. Only products with both high-quality images and detailed textual descriptions are retained to support effective multimodal learning. This rigorous quality assurance process ensures that MMECInstruct provides a reliable and standardized dataset for evaluating MFMs in e-commerce. Details of the dataset processing are in Appendix A.

3.4 Dataset Partitioning

Raw datasets of the CC, PRP, and SA tasks are first split into training, validation, and test data at 8:1:1 ratio. For the AP, PSI, and MPC tasks, the raw datasets are already split. For the SR task, we fol-

Tasks	Training	Validation	IND	OOD
AP, CC, PRP, SA, SR	8,000	1,000	1,000	1,000
PST, MPC	8,000	1,000	1,000	\times
MMECInstruct	56,000	7,000	7,000	5,000

Table 3: Summary of the MMECInstruct dataset. IND and OOD refer to the in-domain evaluation and out-of-domain evaluation, respectively.

low the convention (Hou et al., 2022), leaving the last products in sequence interactions as the test data and the second last products as validation data. Table 3 summarizes the different splits.

Training Set MMECInstruct contains 8K samples for each individual task. These are combined into a single set of 56,000 samples, forming the complete training set for MMECInstruct.

Validation Set MMECInstruct includes a validation set of 1K samples for each individual task. These validation sets are combined into a single set of 7,000 samples, forming the complete validation set for MMECInstruct.

In-domain (IND) Test Set For each of the 7 tasks, MMECInstruct also includes an in-domain test set consisting of 1K samples. IND is defined in terms of products that belong to the same set of categories as those used in the training set.

Out-of-domain (OOD) Test Set To assess the generalizability of models to unseen samples and address the cold-start issue (Schein et al., 2002; Lika et al., 2014) in e-commerce, we create OOD test sets in MMECInstruct. OOD is defined as new products that are not seen during training, identified by their category information. Five tasks (AP, CC, PRP, SA, and SR) have products from various categories. Samples from certain categories are held out as OOD sets. We focus on new products instead of new users because user identifiers are anonymous in the dataset.

3.5 High-quality Instructions

High-quality instructions are critical to the effective adaptation of general-purpose LLMs to e-commerce (Peng et al., 2024; Jin et al., 2024a,b). In MMECInstruct, to ensure its high quality, we carefully craft a instruction for each of the seven e-commerce tasks. Each instruction has been meticulously evaluated and refined by human experts to ensure clarity, conciseness, and accuracy. The detailed description of instructions is in Appendix B.

4 CASLIE: Lightweight Learning Framework for E-commerce MFMs

MMECInstruct presents a multimodal dataset designed to evaluate how well models can effectively leverage both visual and textual information for e-commerce tasks. While directly fine-tuning general multimodal models may seem like a straightforward solution, the results of the fine-tuned MFMs (discussed in Section 6.1) indicate that these models struggle with domain-specific challenges. To address this, we introduce CASLIE, which consists of three key modules: (1) an enriched module (EC³) that generates context-conditioned captions from images (Section 4.1), (2) a light-weighted module (CQE) that evaluates caption qualities (Section 4.2), and (3) a light-weighted multimodal information fusion module that integrates high-quality captions with item context information (Section 4.3) to perform e-commerce tasks. Figure 1b presents an overview of CASLIE. We provide an analysis in Appendix C to explore the impact of captioning models in EC³ and caption quality evaluation models in CQE on the performance of CASLIE.

4.1 Enriched Context-conditioned Captioning

CASLIE first employs a novel enriched context-conditioned captioning module – EC³, to generate textual captions for images, conditioned on the corresponding context, such as user queries or reviews. Unlike CLIP-based models (Chia et al., 2022; Stevens et al., 2024), which implicitly assume that the image in its entirety is relevant to the context. EC³ selectively highlights image details pertinent to the given context.

EC³ utilizes the strong image understanding capability of pre-trained MFMs for conditioned caption generation via zero-shot prompting, integrating context information with well-elaborated instructions to form a prompt (detailed in Appendix B). A unique advantage of using pre-trained MFMs is that their extensive world knowledge, allowing EC³ to enrich captions with relevant insights beyond what is explicitly visible in the images, and thus, benefiting target tasks. We use Llama-3.2-Vision-Instruct (Dubey et al., 2024) as the EC³ model.

4.2 Caption Quality Evaluation

Existing multimodal e-commerce methods use all available images equally (Zhuge et al., 2021; Gao et al., 2020) without evaluating their potential contributions to the target tasks. We denote this strat-

egy as UIA (use it always). However, not all product images are high-quality or contain pertinent information, particularly under different contextual conditions. To ensure that the visual data contributes effectively and meaningfully in different conditions, CASLIE incorporates a caption quality evaluation module – CQE, to assess whether the generated captions, and thus the corresponding product images, meaningfully contribute to the task and should be utilized.

CQE evaluates caption qualities by determining whether or not the captions provide beneficial information for the target task via binary classification. It employs powerful LLMs and MFMs as classifiers, leveraging the contextual information and well-curated instructions (detailed in Appendix B) for zero-shot evaluations, predicting if the generated caption should be utilized. To mitigate inconsistencies in LLM-based predictions (Bonagiri et al., 2024), CQE aggregates outputs from five LLMs via majority voting, denoted as MV, to reach a consensus as the final decision. CASLIE integrates only captions deemed beneficial, enabling a more strategic and deliberate fusion of multimodal data. We use five generalist models as the binary classifiers for MV: Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.2-Vision-Instruct (Dubey et al., 2024), as well as Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Phi-3.5-mini-Instruct (Abdin et al., 2024).

4.3 Modality-unified E-commerce Module

Through EC³ and CQE, CASLIE explicitly translates visual content (i.e., images) into useful textual representations (i.e., captions). These textual representations can be seamlessly integrated with other textual information (e.g., product titles or user reviews) by concatenating them. Such concatenated texts will be used as input and the corresponding response as output to fine-tune a modality-unified e-coMMerCe Module, denoted as uniM³. Three variants with various sizes for uniM³ are fine-tuned: (1) uniM³-L with Llama-2-13B-chat (Touvron et al., 2023), (2) uniM³-M with Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and (3) uniM³-S with Llama-3.2-3B-Instruct (Dubey et al., 2024) as the base models, respectively. These models are optimized using LoRA (Hu et al., 2022) and Huggingface transformers library (Wolf, 2019) on the MMECInstruct dataset. We refer to these models fine-tuned with the CASLIE learning framework as CASLIE-L, CASLIE-M, and CASLIE-S, respectively.

5 Experimental Setup

Baselines We compare CASLIE against 4 categories of baseline methods. (1) fine-tuned MFMs: LLaVA-Interleave (Li et al., 2024a), (2) e-commerce LLMs: eCeLLM-L and eCeLLM-M (Peng et al., 2024), (3) fine-tuned CLIP-based models: FashionCLIP (Chia et al., 2022), and (4) textual task-specific models. More detailed experimental setup is reported in Appendix D. We conduct IND and OOD evaluation (Section 3) for all the methods. The fine-tuned models and textual task-specific models are trained on MMECInstruct. More details on the experimental setup are available in Appendix D.

6 Experimental Results

We conduct a systematic evaluation of CASLIE against all the baselines using the test set of each individual task in MMECInstruct. For a comprehensive evaluation, we utilize multiple metrics on each task. To enable a succinct presentation, for each task, we present only the performance at the primary metric, defined as follows: F1 score for AP and PSI, Recall@1 for CC and SR, accuracy for MPC, macro F1 score for PRP and SA. Complete results for each task are reported in Appendix E. When comparing CASLIE with baselines, we report the mean of CASLIE’s improvement over baselines per task as its overall improvement. Additional results on the in-domain evaluation and complete evaluation results for all the e-commerce tasks are available in E.

6.1 In-domain Evaluation

The left part of Table 4 shows the overall performance in IND evaluation.

(1) CASLIE-M *substantially outperforms the baselines at 6.4% on across 7 tasks* (average of the improvement on each task) as shown in Table 4. These results demonstrate the remarkable effectiveness of CASLIE compared with the fine-tuned CLIP-based model, fine-tuned LLMs, e-commerce LLMs, fine-tuned MFMs, and the task-specific models across the widely-performed e-commerce tasks.

(2) CASLIE-M *achieves a considerable 52.9% improvement over the fine-tuned MFM ft-LLaVA-NExT-Interleave*, as demonstrated in Table 4. Notably, the most significant performance gap occurs on the SR task (0.223 vs. 0.053), which involves processing multiple images. ft-LLaVA-NExT-

Model	IND							OOD				
	AP	CC	PRP	PSI	MPC	SA	SR	AP	CC	PRP	SA	SR
ft-LLaVA-Interleave	0.791	<u>0.964</u>	<u>0.568</u>	<u>0.340</u>	<u>0.721</u>	0.561	0.053	0.579	0.043	0.334	0.206	0.000
eCeLLM-L	0.872	0.870	0.519	0.178	0.706	0.613	<u>0.188</u>	<u>0.860</u>	0.916	0.531	0.584	<u>0.304</u>
eCeLLM-M	0.864	0.890	0.492	0.131	0.719	<u>0.632</u>	0.182	0.841	<u>0.942</u>	<u>0.564</u>	<u>0.624</u>	0.302
ft-FashionCLIP	0.759	0.863	0.497	0.201	0.605	0.323	0.145	0.600	0.903	0.453	0.376	0.087
Task-specific Model	<u>0.868</u>	0.671	0.531	0.316	0.702	0.495	0.163	0.849	0.658	0.447	0.510	0.210
CASLIE-L	0.868	0.969	0.473	0.268	0.706	0.651	0.190	0.840	0.968	0.531	0.607	0.297
CASLIE-M	0.891	0.979	0.566	0.398	0.731	0.656	0.223	0.855	0.977	0.585	0.625	0.330
CASLIE-S	0.871	0.963	0.504	0.336	0.707	0.601	0.196	0.857	0.959	0.580	0.647	0.297
imprv over best (%; avg: 5.3)	2.6	1.6	-0.4	17.1	1.4	3.8	18.6	-0.3	3.7	3.7	3.7	8.6
caption used (%; avg: 45.0)	62.1	62.3	50.5	74.5	72.2	56.8	30.3	68.2	62.6	43.2	56.4	30.4

Table 4: Overall performance comparison. The best performance of CASLIE is in **bold** and of baselines is in underlined. The “imprv over best” refers to the improvement of CASLIE over the best baselines; “caption used” refers to the percentage of captions selected by MV.

Interleave directly encodes raw images alongside text in a fixed interleaved format, treating all visual content indistinguishably regardless of context. On the contrary, CASLIE uses visual content differentially via context-conditioned captioning, emphasizing task-related information from images. This process enables CASLIE to focus on the most informative image content while discarding irrelevant or noisy inputs, leading to significantly better performance, particularly on complex tasks like SR.

(3) CASLIE *exhibits superior performance over e-commerce LLMs*. Specifically, CASLIE-M outperforms eCeLLM-L by 25.2% and eCeLLM-M by 37.1%. The results highlight the benefit of incorporating contextually relevant product image information into CASLIE, while eCeLLM models only utilize textual data.

We provide more analysis on IND evaluation compared to e-commerce LLMs and task-specific models in Appendix E.1, as well as the comparison with proprietary models and the error analysis in Appendix E.2 and E.3. In general, CASLIE-M outperforms both ft-FashionCLIP and task-specific models by 45.8% and 22.1% gains, respectively. Moreover, the mid-size CASLIE-M offers the best performance, benefiting from its powerful base model.

6.2 Out-of-domain Evaluation

The right part of Table 4 presents the performance of CASLIE and baselines in OOD evaluation. Overall, CASLIE *demonstrates strong generalizability to handle products in new categories*, with CASLIE-M outperforming the best baselines by

3.9% average improvement.

(1) CASLIE-M *surpasses the ft-LLaVA-NExT-Interleave by a substantial 624.6% improvement across 4 tasks except for SR in the OOD setting*, underscoring the strong generalizability of CASLIE. ft-LLaVA-NExT-Interleave appears to be struggling to transfer knowledge effectively in OOD scenarios, possibly due to that products from new categories may have very different images. CASLIE takes advantage of the well-known generalizability of LLMs (Touvron et al., 2023; Jiang et al., 2023; Dubey et al., 2024) to understand such new images by translating images to context-conditioned textual representations, and thus generalizes well.

(2) Similarly, CASLIE-M *demonstrates significant advantages over ft-FashionCLIP and eCeLLM-L in the OOD evaluation*, with average 85.1% and 6.4% improvements, respectively. CASLIE could easily leverage LLMs’ generalizability and world knowledge that ft-FashionCLIP doesn’t enjoy. Meanwhile, the ability to integrate visual information via context-conditioned captions allows CASLIE to better capture product details, enabling it to understand new products more effectively than eCeLLM-M, which focuses primarily on text-based information.

6.3 Task-Specific and Generalist CASLIE

Table 5 presents the results of CASLIE fine-tuned with different strategies. When comparing the task-specific CASLIE, which is fine-tuned for each individual task, with the generalist CASLIE, which is fine-tuned across all the tasks together, we observe a trend consistent with that in prior research (Peng et al., 2024): *the generalist CASLIE outperforms*

Setting	IND								OOD				
	AP	CC	PRP	PSI	MPC	SA	SR	AP	CC	PRP	SA	SR	
uniM ³	-L	0.866	0.969	0.468	0.235	0.700	0.628	0.184	0.831	0.959	0.523	0.595	0.285
	-M	0.876	0.971	0.533	0.312	0.725	0.617	0.218	0.847	0.965	0.530	0.659	0.312
	-S	0.866	0.951	0.493	0.270	0.699	0.565	0.191	0.838	0.962	0.511	0.614	0.305
UIA	-L	0.859	0.973	0.486	0.268	0.704	0.607	0.135	0.840	0.968	0.533	0.606	0.236
	-M	0.885	0.976	0.535	0.352	0.722	0.642	0.207	0.859	0.976	0.532	0.613	0.310
	-S	0.869	0.958	0.503	0.299	0.702	0.578	0.196	0.856	0.957	0.515	0.565	0.280
CASLIE	-L	0.868	0.969	0.473	0.268	0.706	0.651	0.190	0.840	0.968	0.531	0.607	0.297
	-M	0.891	0.979	0.566	0.398	0.731	0.656	0.223	0.855	0.977	0.585	0.625	0.330
	-S	0.871	0.963	0.504	0.336	0.707	0.601	0.196	0.857	0.959	0.580	0.647	0.297

Table 6: Ablation study on different module settings of CASLIE of large (-L), middle (-M) or small (-S) sizes. uniM³ is the ablated version that uses text-only input. CASLIE-UIA is the ablated version that always uses the visual information without quality evaluation. The best performance on each task is in **bold**.

Size	Training	AP	CC	PRP	PSI	MPC	SA	SR
-L	T-spec.	0.837	0.931	0.428	0.000	0.671	0.553	0.058
	Gen.	0.868	0.969	0.473	0.205	0.706	0.651	0.190
-M	T-spec.	0.866	0.968	0.495	0.000	0.709	0.600	0.197
	Gen.	0.891	0.979	0.566	0.398	0.731	0.656	0.223
-S	T-spec.	0.838	0.912	0.460	0.000	0.684	0.557	0.121
	Gen.	0.871	0.963	0.504	0.336	0.707	0.601	0.196

Table 5: Comparison of task-specific (T-spec.) and generalist (Gen.) CASLIE models of large (-L), middle (-M) or small (-S) sizes.

task-specific CASLIE on each individual task. Generalist CASLIE-L, CASLIE-M, and CASLIE-S exhibit significant improvements of 44.8%, 7.3%, and 15.4% over their respective task-specific CASLIE across all tasks except for PSI. These results highlight that training on all tasks together, CASLIE enjoys strong versatility and learns transferable knowledge across tasks to boost the performance on individual tasks. It is noteworthy that on PSI, all task-specific CASLIE models fail due to highly unbalanced labels (74% negatives), whereas generalist CASLIE models still achieve considerable performance. This demonstrates that certain e-commerce tasks (e.g., PSI) could substantially benefit from knowledge transfer through generalist modeling.

6.4 Ablation Study

In Table 6, we compare the CASLIE framework with two ablated versions with selected modules: uniM³ uses text-only input, and CASLIE-UIA always uses the visual information without quality evaluation. Take the mid-size models as examples, CASLIE-M **brings a 4.1% average improvement compared to** CASLIE-UIA-M, and a 4.9% average improvement over uniM³-M, highlighting the im-

portance of conditioned captioning and selective visuals integration.

These observations underscore the key benefits of CASLIE’s modular design to integrate selective (by CQE) text-based image representation (by EC³) into uniM³. CASLIE gains benefit from EC³ by extracting context-conditioned captions, effectively translating visual information into textual format for later seamless incorporation. The CQE module further refines this process by filtering out non-beneficial image information, ensuring that only task-relevant visual data is integrated. By concatenating textual and selected visual information and feeding them into powerful uniM³, CASLIE enhances its ability to jointly learn e-commerce tasks from a multimodal perspective, enabling performance that text-only information cannot achieve.

Besides, we also conduct ablation studies on using various captioning models in EC³ and various evaluation strategies in CQE, demonstrating the effectiveness of our design in Appendix C.

7 Conclusion

We develop and open-source a high-quality, multimodal instruction dataset for e-commerce. To our knowledge, MMECInstruct is the first of its kind. We also develop CASLIE, a simple, yet effective framework integrating multimodal information for e-commerce. Leveraging MMECInstruct, we fine-tune the state-of-the-art MFMs (CASLIE series) within CASLIE for e-commerce. Our extensive evaluation of CASLIE models against the most advanced baseline models shows that MMECInstruct enhances CASLIE with advanced capabilities and strong generalizability in e-commerce applications.

8 Limitations

First, while our dataset MMECInstruct undergoes rigorous quality control, there remains a possibility that some samples may still contain noisy or inaccurate information (e.g., mismatch between text and image). This might hinder the performance of the CASLIE that is fine-tuned on this dataset. Second, the LLM-based captioning module EC³ might generate inaccurate or even hallucinated captions in rare occasions, where the captions do not truthfully represent actual objects in the images. This issue might be partially addressed via preference alignment and optimization (Gunjal et al., 2024) to reduce hallucination. Third, CQE can only decide whether or not the captions provide beneficial information within the given context but lacks interpretability to explicitly pinpoint the particular regions/details of the images that are beneficial to the tasks. For future work, we plan to leverage image segmentation techniques (Kirillov et al., 2023) to achieve a more fine-grained evaluation of the images. Fourth, our framework is based on manually-crafted prompt templates, which may be suboptimal in certain cases. For future work, we plan to introduce automatic prompt optimization techniques (Pryzant et al., 2023) to create customized prompts tailored to various e-commerce tasks and use cases.

While it is our aspiration that e-commerce models can enrich users' online experience and enhance users' satisfaction, we also acknowledge that unintended use of e-commerce models might introduce popularity bias (Chen et al., 2023) (e.g., only recommend popular products in the sequential recommendation task) among a large group of users. This issue might be exacerbated when the popular products have more, high-quality image data, and thus bias the image data integration in multimodal e-commerce models. This issue can be mitigated by introducing debiasing algorithms (Wang et al., 2021; Zhang et al., 2021) in the future.

9 Ethics Statement

Our dataset MMECInstruct is constructed all based on public, open-sourced datasets with proper licensing to allow for redistribution and research purposes (Table A1). All the user IDs are fully anonymized, and there is no user profile information (e.g., user names, user address) that could lead to potential disclosure of user privacy.

References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. [Flamingo: a visual language model for few-shot learning](#). *Advances in neural information processing systems*, 35:23716–23736.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474.
- Vamshi Krishna Bonagiri, Sreeram Vennam, Manas Gaur, and Ponnurangam Kumaraguru. 2024. Measuring moral inconsistencies in large language models. *arXiv preprint arXiv:2402.01719*.
- Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. [Bias and debias in recommender system: A survey and future directions](#). *ACM Transactions on Information Systems*, 41(3):1–39.
- Lei Chen, Houwei Chou, Yandi Xia, and Hirokazu Miyake. 2021. [Multimodal item categorization fully based on transformer](#). In *Proceedings of the 4th Workshop on e-Commerce and NLP*, pages 111–115, Online. Association for Computational Linguistics.
- Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. 2022. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*.
- Alfredo Daza, Néstor Daniel González Rueda, Mirelly Sonia Aguilar Sánchez, Wilmer Filomeno Robles Espíritu, and María Elena Chauca Quiñones. 2024. Sentiment analysis on e-commerce product reviews using machine learning and deep learning algorithms: A bibliometric analysis and systematic literature review, challenges and future works. *International Journal of Information Management Data Insights*, 4(2):100267.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

- Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. 2025. [What to align in multimodal contrastive learning?](#) In *The Thirteenth International Conference on Learning Representations*.
- Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Hua-jun Chen. 2024. [Mol-instructions: A large-scale biomolecular instruction dataset for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.
- Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.
- Shifan Gu and Yancong Deng. 2024. Exploring the influence of image features and contextual factors on purchase decisions: An empirical analysis in e-commerce. In *Proceedings of the 2024 15th International Conference on E-business, Management and Economics*, pages 229–236.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.
- Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary Chase Lipton. 2019. [AmazonQA: A review-based question answering task](#). In *International Joint Conference on Artificial Intelligence*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.
- Bethany L Hedt-Gauthier, Tisha Mitsunaga, Lauren Hund, Casey Olives, and Marcello Pagano. 2013. The effect of clustering on lot quality assurance sampling: a probabilistic model to calculate sample sizes for quality assessments. *Emerging themes in epidemiology*, 10:1–11.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.
- Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 585–593.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, et al. 2024a. Amazonm2: A multilingual multi-locale shopping session dataset for recommendation and text generation. *Advances in Neural Information Processing Systems*, 36.
- Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. 2023. [Learning instance-level representation for large-scale multi-modal pretraining in e-commerce](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11069.
- Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, et al. 2024b. Shopping mmlu: A massive multi-task online shopping benchmark for large language models. *arXiv preprint arXiv:2410.20745*.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. [Segment anything](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

- Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8188–8195.
- Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020b. Keywords-guided abstractive sentence summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8196–8203.
- Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023a. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1258–1267.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hoi. 2021. [Align before fuse: Vision and language representation learning with momentum distillation](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705.
- Xin Li, Mengyue Wang, and Yubo Chen. 2014. The impact of product photo on online consumer purchase intention: An image-processing enabled empirical study. In *18th Pacific Asia Conference on Information Systems, PACIS 2014*, page 325. Association for Information Systems.
- Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2024b. EcomGPT: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18582–18590.
- Yunxin Li, Baotian Hu, Wenhan Luo, Lin Ma, Yuxin Ding, and Min Zhang. 2024c. A multi-modal in-context tuning approach for e-commerce product description generation. *arXiv preprint arXiv:2402.13587*.
- Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert systems with applications*, 41(4):2065–2073.
- Xixun Lin, Rui Liu, Yanan Cao, Lixin Zou, Qian Li, Yongxuan Wu, Yang Liu, Dawei Yin, and Guandong Xu. 2025. Contrastive modality-disentangled learning for multimodal recommendation. *ACM Transactions on Information Systems*, 43(3):1–31.
- Fan Liu, Delong Chen, Xiaoyu Du, Ruizhuo Gao, and Feng Xu. 2023a. Mep-3m: A large-scale multimodal e-commerce product dataset. *Pattern Recognition*, 140:109519.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). *Advances in neural information processing systems*, 36.
- Qidong Liu, Jiayi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2023c. Multimodal recommender systems: A survey. *ACM Computing Surveys*.
- Yao Mu, Junting Chen, Qinglong Zhang, Shoufa Chen, Qiaojun Yu, Chongjian Ge, Runjian Chen, Zhixuan Liang, Mengkang Hu, Chaofan Tao, et al. 2024. Robocodex: Multimodal code generation for robotic behavior synthesis. *arXiv preprint arXiv:2402.16117*.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.
- 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](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 40215–40257. PMLR.
- Bo Peng, Srinivasan Parthasarathy, and Xia Ning. 2023. Multi-modality meets re-learning: Mitigating negative transfer in sequential recommendation. *arXiv preprint arXiv:2309.10195*.
- Aleksandr Vladimirovich Petrov and Craig Macdonald. 2023. gsa-rec: reducing overconfidence in sequential recommendation trained with negative sampling. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 116–128.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. [Automatic prompt optimization with “gradient descent” and beam search](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968.

- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- 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. *arXiv preprint arXiv:2206.06588*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260.
- Kaize Shi, Xueyao Sun, Dingxian Wang, Yinlin Fu, Guandong Xu, and Qing Li. 2023. Llama-e: Empowering e-commerce authoring with multi-aspect instruction following. *arXiv preprint arXiv:2308.04913*.
- Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. 2024. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Jinpeng Wang, Ziyun Zeng, Yunxiao Wang, Yuting Wang, Xingyu Lu, Tianxiang Li, Jun Yuan, Rui Zhang, Hai-Tao Zheng, and Shu-Tao Xia. 2023. [MISSRec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation](#). In *Proceedings of the 31st ACM International Conference on Multimedia, MM '23*, page 6548–6557, New York, NY, USA.
- Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. [Deconfounded recommendation for alleviating bias amplification](#). In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1717–1725.
- Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- T Wolf. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020a. Product knowledge graph embedding for e-commerce. In *Proceedings of the 13th international conference on web search and data mining*, pages 672–680.
- Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020b. Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1355–1362.
- Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1256–1265.
- Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L Berg, and Ning Zhang. 2022. [CommerceMM: Large-scale commerce multimodal representation learning with omni retrieval](#). In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4433–4442.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. In *The Twelfth International Conference on Learning Representations*.
- Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. [Causal intervention for leveraging popularity bias in recommendation](#). In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 11–20.
- Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. *arXiv preprint arXiv:2009.07162*.
- Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-bert: Vision-language

pre-training on fashion domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12647–12657.

A Dataset Details

To pursue adherence to data usage requirements, we check the licenses of MMECInstruct data sources, ensuring their permission to publish. Table A1 presents the licenses of our curated dataset sources.

A.1 Task Selection

Following ECInstruct (Peng et al., 2024), MMECInstruct comprises 7 widely-performed real-world tasks constructed from real-world data, which are ubiquitous and essential in the e-commerce domain elaborating in Table 2. Not all ECInstruct tasks are involved since some data sources lack vision information. Previous methods for summarization (Xu et al., 2020b; Li et al., 2020b,a), extraction (Zhu et al., 2020) and description generation (Li et al., 2024c) also aim for generation tasks in e-commerce domain but study different direction from this work. Therefore these tasks are not considered here. Following prior research (Wei et al., 2022) and taking into account the high computational demands, we uniformly downsample the training sets for each individual task to 8K samples, the validation sets to 1K, and the test sets to 1K. This ensures an optimal balance between data volume and efficient processing for affordable LLM evaluation.

A.2 Data Selection

In the AP, PRP, SA, and SR tasks, Tools category data from Amazon datasets (Gupta et al., 2019; Hou et al., 2024; Ni et al., 2019) serve as in-domain (IND) data sources, and Sports category data serves as out-of-domain (OOD) data.

For the MPC and PSI tasks, we directly process the row datasets (Reddy et al., 2022) from their original splits.

For the CC tasks, we select the 100 most frequent fine-grained categories as in-domain (IND) data, while categories ranked between 100 and 200 in frequency are used as out-of-domain (OOD) data.

A.3 Data Statistics

Figure A1 presents the distributions of input lengths for each task, measured by word count. For better clarity, we exclude very long inputs (those representing at most 1% of samples) in the SA and SR tasks.

Table A2 presents the distribution of product categories in the MMECInstruct dataset. The dataset

spans a wide variety of categories, reflecting the heterogeneity of real-world e-commerce platforms. Notably, it includes high-volume categories and also incorporates lower-frequency and long-tail categories, enhancing its diversity. This stratified coverage across both popular and niche domains enables MMECInstruct to support robust training and evaluation of multimodal models under varied product scenarios.

A.4 Data Processing

We conduct the data processing following ECInstruct (Peng et al., 2024) as below. Besides that, we thoroughly check the availability of each product’s image.

A.5 Dataset Partitioning

Answerability Prediction (AP) We utilize the data from the Tools category of AmazonQA (Gupta et al., 2019) as the in-domain (IND) source and the Sports category as the out-of-domain (OOD) source for this task. The *is_answerable* annotations serve as the ground truth. In the structured dataset, the ratio of positive to negative samples is approximately 3:5.

Category Classification (CC) We use the fine-grained product category labels from MAVE (Yang et al., 2022) as the ground truth. To ensure each selected category has sufficient data, we first sort the categories by frequency. We then select the 100 most frequent fine-grained categories as IND data, while categories ranked between 100 and 200 in frequency are designated as OOD data. Then we split IND data with an 8:1:1 ratio to formulate training, validation, and IND test set.

Product Relation Prediction (PRP) Similar to ECInstruct (Peng et al., 2024), to study product relationships, we utilize the product metadata from the Tools category as IND sources, with the Sports category serving as the OOD source. We collect product IDs from the metadata, removing any products that lack detailed information. Product titles and images are used to represent the products in this task, and any product pairs that appear multiple times with different relations are eliminated. After filtering and integrating the data with instruction templates, the three types of relationships (*also buy*, *also view*, and *similar*) are distributed in the final dataset at approximately a 4:3:1 ratio.

Dataset	License Type	Source
Amazon Review	Not Specified	https://https://amazon-reviews-2023.github.io/
AmazonQA	Not Specified	https://github.com/amazonqa/amazonqa
MAVE	CC-by-4.0	https://github.com/google-research-datasets/MAVE
Shopping Queries Dataset	Apache License 2.0	https://github.com/amazon-science/esci-data

Table A1: Details of Data Source License

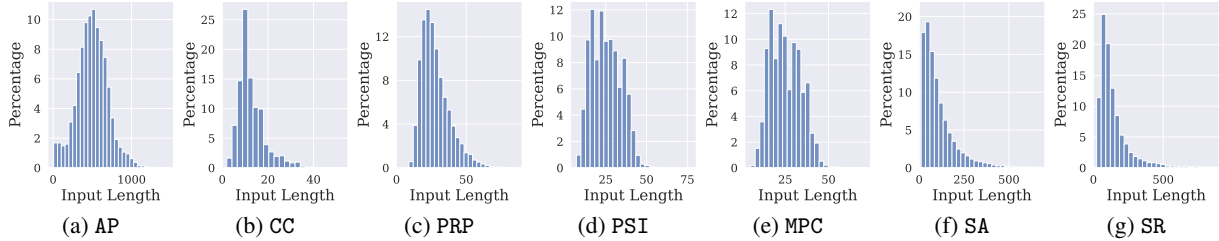


Figure A1: Distribution of Input Length in MMECInstruct

Category	Percentage (%)
Cell Phones and Accessories	18.91
Tools and Home Improvement	14.82
Electronics	11.96
Home and Kitchen	11.02
Clothing Shoes and Jewelry	6.00
Sports and Outdoors	5.81
Toys and Games	4.65
Books	3.64
Automotive	2.83
Beauty and Personal Care	2.81
Grocery and Gourmet Food	2.75
Health and Household	2.68
Patio Lawn and Garden	2.06
Office Products	1.93
Arts Crafts and Sewing	1.48
Pet Supplies	1.47
Others	5.18

Table A2: Category Statistics

Product Substitute Identification (PSI) We represent products from the Shopping Queries dataset (Reddy et al., 2022) using their titles and images and eliminate non-English samples. Each query-product pair is labeled into 4 categories (*Exact*, *Substitute*, *Complement*, and *Irrelevant*) The query-product pairs with *Exact*, *Complement*, or *Irrelevant* labels are relabeled as non-substitute. The ratio of the positive and negative labels in the MMECInstruct dataset is approximately 1:3.

Multi-class Product Classification (MPC) The preprocessing of the MPC is similar to that of PSI, except that the MPC is a multi-class classification task. The ratio of the four labels in the structured dataset (*Exact*, *Substitute*, *Complement*, and *Irrelevant*) is approximately 20:7:1:4.

Sentiment Analysis (SA) For the sentiment analysis, we use the review data of the Tools category from the Amazon Review dataset (Hou et al., 2024) as the IND sources and the Sports category as the OOD source. We only retain the reviews that are longer than 10 words.

Sequential Recommendation (SR) In the SR task, we utilize both product reviews and meta-data from the Amazon Review dataset (Hou et al., 2024). Additionally, we incorporate users’ review histories as a representation of their interactions with products. The processing protocol follows the same steps as ECInstruct (Peng et al., 2024), with the primary distinction being the inclusion of images for each product. The curated dataset has an average of 10.7 interactions per user and an average text length of 18 words per product.

B Instruction Templates

B.1 Answerability Prediction (AP)

Captioning Instruction Please generate an informative caption for the product in the image. The caption should be helpful to identify if the product-related question: {{question}}, is answerable.

Caption Quality Evaluation Instruction The task needs to identify if the question is answerable based on the related document: {{review}}. Here is the additional information about the product that was extracted from the product image: {{caption}}. You need to determine if the information extracted from the image will help to identify the question’s answerability. Only output yes or no.

Task Instruction Analyze the question and its supporting document, as well as the potential extra information about the products extracted from the product images, predict if the question is answerable based on the provided information. Output only yes or no.

B.2 Category Classification (CC)

Captioning Instruction Please generate an informative caption for the product in the image. Here is the product title: {{title}}. The caption should be helpful in identifying the product's fine-grained category.

Caption Quality Evaluation Instruction The task needs to identify the product's fine-grained category from the options: {{options}}. Here is the additional information about the product that was extracted from the product image: {{caption}}. You need to determine if the information extracted from the image will help to identify the category. Only output yes or no.

Task Instruction Analyze the product title, as well as the potential extra information about the products extracted from the product images, identify the product category from the given options. Only answer from the options.

B.3 Product Relation Prediction (PRP)

Captioning Instruction Please generate an informative caption for the product in the image. The title of the product in the image is {{title of the product}}. The caption should be helpful in predicting the relation between this product and {{title of another product}}.

Caption Quality Evaluation Instruction The model needs to identify if the two products are similar or will be purchased together or be viewed together given the title of product 1: {{title of the product}}, and product 2: {{title of another product}}. Here is the additional information about product 1 extracted from its image: {{caption of product 1}}, you need to determine if the information extracted from the image will be helpful in identifying the relation between the two products. Only output yes or no.

Task Instruction Given the title of two products, as well as the potential extra information about the products extracted from the product images, predict the relation of the two products. Only answer from the options.

B.4 Product Substitute Identification (PSI)

Captioning Instruction Please generate an informative caption for the product in the image. The caption should be helpful to predict if the product: {{title}} can serve as a functional substitute for the user's query: {{query}}.

Caption Quality Evaluation Instruction The model needs to identify if the product is somewhat relevant to the query but fails to fulfill some aspects of the query but the product can be used as a functional substitute. Given a user's query: {{query}} and a product title: {{title}}, as well as additional information about the product extracted from the product image: {{caption}}, you need to determine if the information extracted from the image will be helpful in identifying the relevance between the product and the query. Only output yes or no.

Task Instruction Given a user's query and a product title, as well as the potential extra information about the product extracted from the product image, identify if the product is somewhat relevant to the query but fails to fulfill some aspects of the query but the product can be used as a functional substitute. Only output yes or no.

B.5 Multi-class Product Classification (MPC)

Captioning Instruction Please generate an informative caption for the product in the image. The caption should be helpful to predict the relevance between the user's query: {{query}}, and product: {{title}}.

Caption Quality Evaluation Instruction The model needs to predict the relevance between the query and product by analyzing the user's query: {{query}}, and product title: {{title}}. Here is the additional information about the product extracted from the product image: {{caption}}, you need to determine if the information extracted from the image will be helpful in predicting the result. Only output yes or no.

Task Instruction Predict the relevance between the query and product by analyzing the user's query, and product title, as well as the potential extra information about the product extracted from the product image. Output the option that best describes the relevance.

B.6 Sentiment Analysis (SA)

Captioning Instruction Please generate an informative caption for the product in the image. The

caption should be helpful to identify the user’s sentiment from the review: {{review}}.

Caption Quality Evaluation Instruction The task needs to identify the user’s sentiment based on their review: {{review}}. Here is the additional information about the product extracted from the user review’s image: {{caption}}. You need to determine if the information extracted from the image will help to identify the user’s sentiment. Only output yes or no.

Task Instruction Given the user’s review, as well as the potential extra information about the products extracted from the user review’s image, identify the user’s sentiment. Only answer from the options.

B.7 Sequential Recommendation (SR)

Captioning Instruction Please generate an informative caption for the product in the image. Here is the product title: {{title}}. The caption should be helpful in predicting the next product the user is most likely to purchase by analyzing the user’s intent based on the user’s purchase history.

Caption Quality Evaluation Instruction The task needs to recommend the next product that the user may be interested in based on the user’s purchase history. Here is the title of a product from purchase history: {{title, category, brand}}, and the information extracted from the product image: {{caption}}. You need to determine if the information extracted from the image will be helpful for recommendation. Only output yes or no.

Task Instruction Estimate the user’s intent based on the user’s purchase history, and predict the next product that the user is most likely to purchase from the given options.

C Analysis on EC³ and CQE

In this section, we explore the impact of captioning models in EC³ and caption quality evaluation models in CQE on the performance of CASLIE, exemplified by CASLIE-M.

C.1 Analysis on Captioning Models

When analyzing the impact of captioning models, we include BLIP2-OPT-2.7B (Li et al., 2023b) as a context-free captioning model and evaluate it as a baseline. Table A3 also compares the CASLIE-M using various individual captioning models, including

LLaVA-1.5-7B (Liu et al., 2023b, 2024a), LLaVA-NExT-mistral-7B (Liu et al., 2024b), and Llama-3.2-Vision-Instruct (Dubey et al., 2024). Table A3 presents the results.

(1) *Overall, using visual information through captioning is almost always better than not using visual information.* Specifically, using BLIP2-OPT-2.7B to generate context-free captions from images brings a 1.8% average improvement compared with uniM³-M, which does not use visual information at all; using LLaVA-NExT-mistral-7B in CASLIE for context-conditioned captioning results in 8.6% improvement over uniM³-M. This shows the utility of visual information in e-commerce tasks and demonstrates that captioning is an effective way of utilizing images in e-commerce models.

(2) *Context-condition captioning beats context-free captioning for e-commerce.* CASLIE-M, which employs Llama-3.2-Vision-Instruct as the captioning model by default, outperforms that using the context-free captioning model (BLIP2-OPT-2.7B) by 4.5%. This further highlights the advantage of using context-conditioned captioning to enhance task performance compared to more generic, context-free approaches. Comparing all context-conditioned captioning models, we observe comparable results, but *Llama-3.2-Vision-Instruct as the captioning model is slightly and consistently better overall.*

(3) *CASLIE is possessed with better capability leveraging captions than MFM.* ft-LLaVA-NExT-Interleave using captions for the text input improves AP and PSI slightly compared to its image-using counterpart. However, this approach falls behind CASLIE-M across most tasks. This indicates that using captions as a substitute for the original multimodal input in MFMs is suboptimal. MFMs are designed to process multimodal inputs directly, leveraging both visual and textual modalities simultaneously, and are not fully optimized for text-only inputs. The results underscore that simply incorporating captions into MFMs is insufficient to fully leverage the multimodal information cohesively and effectively.

C.2 Analysis on Evaluation Strategies

In Table A4, we compare CASLIE-M using different caption quality evaluation strategies, including using a single evaluation model, and majority voting (MV) from 3, 5, and 7 models. For majority voting with 3 CQE models, we use Llama-3.1-8B-instruct, Llama-3.2-vision-instruct, and Mistral-7B-instruct-

Model	Setting	Captioning Model	AP	CC	PRP	PSI	MPC	SA	SR
			F1	R@1	M-F1	F1	Acc	M-F1	R@1
ft-LLaVA-NExT-Interleave	w image	-	0.791	0.964	0.568	0.340	0.721	0.561	0.053
ft-LLaVA-NExT-Interleave*	w caption	Llama-3.2-Vision-Instruct	0.633	0.961	0.552	0.404	0.722	0.579	0.000
uniM ³ -M	w/o caption	-	0.876	0.971	0.533	0.312	0.725	0.617	0.218
CASLIE-M	w/o context	BLIP2-OPT-2.7B	0.878	0.976	0.545	0.352	0.734	0.614	0.209
		Llama-3.2-Vision-Instruct	0.880	0.978	0.520	0.392	0.727	0.633	0.214
	w/ context & caption	LLaVA-1.5-7B	0.886	0.987	0.532	0.450	0.725	0.637	0.213
		LLaVA-NExT-mistral-7B	0.886	0.979	0.558	0.476	0.725	0.647	0.210
		Llama-3.2-Vision-Instruct	0.891	0.979	0.566	0.398	0.731	0.656	0.223

Table A3: Comparison using Different Captioning Models. The best performance on each task is in **bold**. When employing different caption models, we only involve captions that are predicted to be useful by CQE. * indicates the version of LLaVA-NExT-Interleave fine-tuned and evaluated on captioning data generated by EC³ and CQE.

Strategy	Evaluation Model	AP	CC	PRP	PSI	MPC	SA	SR
		F1	R@1	M-F1	F1	Acc	M-F1	R@1
UIA	-	0.885	0.976	0.535	0.352	0.722	0.642	0.207
Single	Llama-3.2-3B-Instruct	0.884	0.971	0.512	0.395	0.731	0.603	0.216
	Phi-3.5-mini-Instruct	0.885	0.976	0.515	0.294	0.733	0.638	0.210
	Mistral-7B-Instruct-v0.3	0.879	0.976	0.540	0.389	0.737	0.651	0.212
	Llama-3.1-8B-Instruct	0.885	0.974	0.549	0.404	0.722	0.622	0.220
	Llama-3.2-Vision-Instruct	0.885	0.969	0.538	0.397	0.737	0.622	0.223
MV	3 models	0.881	0.969	0.543	0.396	0.719	0.631	0.218
	5 models	0.891	0.979	0.566	0.398	0.731	0.656	0.223
	7 models	0.882	0.984	0.546	0.416	0.740	0.659	0.219

Table A4: Comparison of Caption Quality Evaluation Methods in IND Evaluation. The best performance on each task is in **bold**. The results are evaluated from CASLIE-M.

v0.3 as evaluation models. For five-model voting, we added Phi-3.5-mini-instruct and Llama-3.2-3B-instruct as evaluation models. For seven-model voting, we further include Llama-3-8B-instruct and qwen2.5-7B-instruct as evaluation models. We also compare the strategy when the caption is used always (i.e., UIA), all with Llama-3.2-Vision-Instruct serving as the captioning model (EC³).

(1) *Compared with UIA, using caption quality evaluation models brings performance improvement in general.* As shown in Table A4, compared to UIA, using all evaluation models together with MV leads to a considerable average improvement of 4.4%.

(2) *Compared to using a single evaluation model, MV-based evaluation leads to further improvement.* Notably, employing MV-based evaluation, which combines the results of all evaluation models, yields higher performance than using a single evaluation model (1.7% improvement over CASLIE-M with Llama-3.2-Vision-Instruct as the evaluation model) highlighting the effectiveness of our MV evaluation strategy.

(3) *Compared to using a various number of evaluation models by MV, five evaluation models yield the comparable high performance with less cost.* Specifically, incorporating five evaluation models yields a 2.1% average improvement compared to three models. However, increasing to seven evaluation models provides only a marginal 0.1% improvement over five models. To balance computational cost and performance, we opted to use five models in the CQE module. The results offer deeper insights into the framework’s design choices and substantiate our approach.

C.3 Analysis on Context-conditioned Captions

While some overlap is natural since both captions and titles describe product attributes, our EC³ module generates context-conditioned captions that go beyond static title information. Unlike titles, which are often short, seller-centric, and lack contextual adaptation, EC³ enriches captions with task-relevant visual evidence. For example, in Figure A5, given a user query highlighting “wings”,

EC³ produces “A Labrador Retriever dressed as a yellow angel with moving wings, designed as a tree topper”, which captures fine-grained, query-relevant visual details absent in the product title.

To quantify overlap, we conducted a systematic analysis of generated captions and product titles. We calculate the Jaccard similarity, which computes the percentage of word overlap between two sentences, and the semantic similarity, which calculates the cosine similarity of two sentences’ embedding. The results are demonstrated in Table A5.

The very low Jaccard similarity scores confirm limited word-level overlap, while the higher semantic similarity reflects that both describe the same product but from complementary perspectives. Crucially, captions highlight visual grounding (e.g., colors, arrangements, subtle details) that titles do not encode. Empirically, our ablations (Table 6, uniM³ vs. CASLIE) demonstrate that unimodal fine-tuning on titles alone cannot match CASLIE’s performance, validating that captions provide distinctive and non-redundant contributions.

C.4 Real-world Considerations

When considering the real-world situation, scalability, computational costs, or integration in environments are important. CASLIE is inherently deployable, as it avoids joint end-to-end multimodal training. Take the MPC task as an example, we calculate the runtime of CQE with 5 LLMs and result in 0.4s per instance since each model only needs to answer yes-no questions. Besides, CASLIE’s modularity allows seamless substitution or refinement of components in e-commerce environments.

D Detailed Experimental Setup

Fine-tuned CLIP-based Models FashionCLIP (Chia et al., 2022) is a SoTA CLIP-based (Radford et al., 2021) model adapted to the e-commerce fashion domain and is skilled at various multimodal tasks. We fine-tune the Huggingface checkpoint of FashionCLIP on each task using the MMECInstruct training set and denoted the fine-tuned model as ft-FashionCLIP.

Fine-tuned MFMs We fine-tune LLaVA-NExT-interleave-qwen-7b (Li et al., 2024a) as the MFM baseline, which is a SoTA multi-image MFM able to process input textual and image information of one or multiple instances, making it a suitable baseline for e-commerce tasks, particularly those evaluating multiple products simultaneously (e.g., PRP).

We fine-tune the checkpoint of LLaVA-NExT-interleave-qwen-7b released in Huggingface on the training data of MMECInstruct. The fine-tuned model is denoted as ft-LLaVA-NExT-interleave. We also conduct the zero-shot evaluation for this baseline.

E-commerce LLMs We utilize eCeLLM-L and eCeLLM-M (Peng et al., 2024), a series of SoTA e-commerce LLMs, fine-tuned on various e-commerce tasks, as a baseline. For eCeLLM-L and eCeLLM-M, we perform a zero-shot evaluation using the checkpoints available on Huggingface since they already encompass a broad understanding of e-commerce concepts.

SoTA Task-Specific Models To evaluate the SR and CC tasks, we fine-tune Recformer (Li et al., 2023a), a popular language-based recommendation model, and Sentence-BERT (Reimers and Gurevych, 2019), which is adept at semantic similarity search tasks like retrieval, respectively. All other tasks are evaluated on the fine-tuned DeBERTa (He et al., 2021), which is a widely used BERT-based model known for its strong performance in various language understanding tasks.

Hyperparameters and Reproducibility The learning rate and batch size are set as 1e-4 and 128 during fine-tuning of all the models. A cosine learning rate scheduler with a 5% warm-up period for 3 epochs is applied. We set α and the rank in LoRA as 16, and add LoRA adaptors to all the projection layers and the language modeling head. We perform zero-shot evaluations (i.e., without in-context examples) on all the tasks.

E Detailed Experimental Results

E.1 More IND Results

In this section, we bring more discussion on in-domain (IND) evaluation as a supplementary of Section 6.1 from Table 4.

(1) CASLIE-M *achieves a significant 45.8% improvement over the ft-FashionCLIP fine-tuned on the training data of MMECInstruct*. A key difference between CASLIE and FashionCLIP is that CASLIE uses the textual representation of images generated via context-conditioned captioning (EC³), adjusting the focus on image details with respect to the specific context. In contrast, FashionCLIP generates image representations without considering the specific context. Additionally, CASLIE could

Similarity	AP	CC	PRP	PSI	MPC	SA	SR
Jaccard Similarity (%)	6.51	17.00	15.51	15.84	16.67	14.36	15.75
Semantic Similarity (%)	46.36	75.89	77.35	70.91	73.17	57.65	75.38

Table A5: Caption-title Similarity

leverage the extensive world knowledge of LLMs to enrich the captions, while FashionCLIP considers the images solely using the vision encoder.

(2) **CASLIE-M outperforms SoTA task-specific models with a significant 22.1% improvement across all 7 tasks.** Compared with SoTA task-specific models, which solely rely on textual information from each individual task, CASLIE could leverage both vision and language information of each task, and the information shared across diverse e-commerce tasks, as well as LLM’s inherent knowledge and learning power, to significantly boost performance on each individual task.

(3) **Mid-size CASLIE-M performs best among CASLIE model sizes.** Benefitting from the large-scale instruction-tuning dataset and powerful base model (Mistral-7B-Instruct-v0.3) mid-size fine-tuned models achieve most, balancing learning from instruction tuning while retaining knowledge from base models.

(4) Considering the percentage of captions selected by MV, **sparse caption usage still leads to high gains, implying a strong signal when captions are selected.** For example, SR only uses captions 30% of the time but leads an 18.6% gain in IND evaluation.

E.2 Comparison with Proprietary Models

We have conducted new experiments with Claude-3.5 and GPT-4o (both text-only and multimodal) to MMECInstruct against our proposed CASLIE models (-S, -M, -L). Evaluation results on IND and OOD test sets are summarized in Table A6.

As shown, CASLIE-M consistently outperforms both Claude-3.5 and GPT-4o across nearly all tasks. Under the IND setting, CASLIE-M achieves the highest overall performance, with particularly large margins on PRP (0.566) and SR (0.223), surpassing GPT-4o (0.441 PRP, 0.123 SR) and Claude-3.5 (0.360 PRP, 0.069 SR). This trend remains consistent in the OOD setting, where CASLIE-M obtains strong generalization. These improvements are particularly pronounced on complex reasoning tasks which require nuanced understanding of contextual and causal relationships.

Furthermore, other variants (CASLIE-S and CASLIE-L) also exhibit competitive or superior performance to both baselines in most metrics, demonstrating the robustness and scalability of the CASLIE architecture. Overall, these results highlight CASLIE’s competitiveness against advanced proprietary models, affirming its strong adaptability and reasoning ability across diverse visual-linguistic domains.

E.3 Error Analysis

We conduct an error analysis with both taxonomy and quantification in using the captions as the visual representation in CASLIE-M by sampling 100 failure cases. The observed errors are categorized into five error types:

(1) **Attribute missing** (18%): image provides a specific attribute, but the caption fails to capture it.

(2) **Attribute hallucination** (7%): caption introduces attributes not grounded in the image.

(3) **Context conflict** (31%): useful product information is diluted or distracted by noisy visual details.

(4) **Helpful caption missing** (10%): beneficial captions are incorrectly filtered out by CQE.

(5) **Hard cases** (34%): captions are accurate, but the task itself is inherently difficult.

Across tasks, we find that context conflict and hard cases dominate. This taxonomy not only clarifies CASLIE’s failure modes but also points to actionable directions: refining caption prompts to reduce missing attributes, improving CQE filtering to recover helpful captions, and exploring debiasing strategies to mitigate context conflicts.

E.4 Detailed Results for All the Tasks

Table A7, A8, A9, A10, A11, A12 and A13 present the complete results for AP, CC, PRP, PSI, MPC, SA and SR, respectively, in IND and OOD evaluation. As shown in these tables, overall, CASLIE models outperform the fine-tuned CLIP-based model (i.e., FashionCLIP), Fine-tuned LLMs (e.g., *ft*-Llama-2-13B), E-commerce LLMs (e.g., eCeLLM-L), the Fine-tuned MFM (i.e., *ft*-LLaVA-NEXT-interleave) and SoTA Task Specific Models in IND evaluation.

Model	IND							OOD				
	AP	CC	PRP	PSI	MPC	SA	SR	AP	CC	PRP	SA	SR
Claude-3.5 (text-only)	0.656	0.949	0.329	0.360	0.683	0.433	0.107	0.659	0.974	0.259	0.427	0.196
Claude-3.5	0.755	0.952	0.360	0.366	0.657	0.481	0.069	0.752	0.976	0.255	0.470	0.139
GPT-4o (text-only)	0.475	0.971	0.401	0.315	0.581	0.525	0.151	0.487	0.989	0.319	0.539	0.213
GPT-4o	0.510	0.982	0.441	0.252	0.664	0.538	0.123	0.551	0.987	0.331	0.565	0.217
CASLIE-L	0.868	0.969	0.473	0.268	0.706	0.651	0.190	0.840	0.968	0.531	0.607	0.297
CASLIE-M	0.891	0.979	0.566	0.398	0.731	0.656	0.223	0.855	0.977	0.585	0.625	0.330
CASLIE-S	0.871	0.963	0.504	0.336	0.707	0.601	0.196	0.857	0.959	0.580	0.647	0.297

Table A6: Performance Comparison with Proprietary Models. The best performance on each task is in **bold**

Model	IND					OOD					
	Acc	M-Rec	M-Pre	M-F1	#Failed	Acc	M-Rec	M-Pre	M-F1	#Failed	
<i>ft</i> -LLaVA-NExT-Interleave	0.746	0.895	0.709	0.791	11	0.509	0.626	0.538	0.579	13	
eCeLLM-L	0.821	0.851	0.894	0.872	0	0.814	0.813	0.912	0.860	0	
eCeLLM-M	0.817	0.876	0.852	0.864	0	0.793	0.809	0.877	0.841	0	
<i>ft</i> -FashionCLIP	0.673	0.764	0.754	0.759	0	0.550	0.677	0.538	0.600	0	
Task-specific Model	0.832	0.939	0.806	0.868	0	0.824	0.917	0.791	0.849	0	
CASLIE-L	uniM ³	0.809	0.832	0.902	0.866	0	0.767	0.760	0.917	0.831	0
	EC ³ -uniM ³	0.799	0.823	0.899	0.859	0	0.781	0.773	0.920	0.840	0
	EC ³ -CQE-uniM ³	0.812	0.833	0.906	0.868	0	0.782	0.776	0.915	0.840	0
CASLIE-M	uniM ³	0.823	0.837	0.919	0.876	0	0.795	0.795	0.906	0.847	0
	EC ³ -uniM ³	0.840	0.866	0.906	0.885	0	0.815	0.820	0.903	0.859	0
	EC ³ -CQE-uniM ³	0.846	0.863	0.921	0.891	0	0.813	0.831	0.880	0.855	0
CASLIE-S	uniM ³	0.808	0.825	0.912	0.866	0	0.772	0.756	0.939	0.838	0
	CASLIE-S-UIA	0.815	0.838	0.903	0.869	0	0.806	0.798	0.923	0.856	0
	CASLIE-S-MV	0.814	0.826	0.921	0.871	0	0.803	0.785	0.944	0.857	0

Table A7: Performance comparison on the AP task. The best performance on each task is in **bold**.

Model	IND		OOD		
	HR@1	#Failed	HR@1	#Failed	
<i>ft</i> -LLaVA-NExT-Interleave	0.964	2	0.043	2	
eCeLLM-L	0.870	0	0.916	0	
eCeLLM-M	0.890	0	0.942	0	
<i>ft</i> -FashionCLIP	0.863	0	0.903	0	
Task-specific Model	0.671	0	0.658	0	
CASLIE-L	uniM ³	0.969	0	0.959	0
	EC ³ -uniM ³	0.973	0	0.968	0
	EC ³ -CQE-uniM ³	0.969	0	0.968	0
CASLIE-M	uniM ³	0.971	0	0.965	0
	EC ³ -uniM ³	0.976	0	0.976	0
	EC ³ -CQE-uniM ³	0.979	0	0.977	0
CASLIE-S	uniM ³	0.951	0	0.962	0
	EC ³ -uniM ³	0.958	0	0.957	0
	EC ³ -CQE-uniM ³	0.963	0	0.959	0

Table A8: Performance comparison on the CC task. The best performance on each task is in **bold**.

Model	IND					OOD					
	Acc	M-Pre	M-Rec	M-F1	#Failed	Acc	M-Rec	M-Pre	M-F1	#Failed	
<i>ft</i> -LLaVA-NExT-Interleave	0.708	0.590	0.570	0.568	6	0.486	0.343	0.326	0.334	6	
eCeLLM-L	0.671	0.654	0.527	0.519	0	0.793	0.534	0.532	0.531	0	
eCeLLM-M	0.690	0.476	0.529	0.492	0	0.843	0.563	0.565	0.564	0	
<i>ft</i> -FashionCLIP	0.630	0.516	0.501	0.497	0	0.622	0.462	0.582	0.453	0	
Task-specific Model	0.704	0.701	0.548	0.531	0	0.665	0.461	0.446	0.447	0	
CASLIE-L	uniM ³	0.659	0.441	0.501	0.468	0	0.782	0.522	0.525	0.523	0
	EC ³ -uniM ³	0.670	0.782	0.514	0.486	0	0.796	0.532	0.534	0.533	0
	EC ³ -CQE-uniM ³	0.666	0.447	0.507	0.473	0	0.692	0.649	0.542	0.531	0
CASLIE-M	uniM ³	0.707	0.666	0.550	0.533	0	0.791	0.533	0.531	0.530	0
	EC ³ -uniM ³	0.705	0.659	0.549	0.535	0	0.793	0.535	0.532	0.532	0
	EC ³ -CQE-uniM ³	0.714	0.708	0.568	0.566	0	0.821	0.610	0.570	0.585	0
CASLIE-S	uniM ³	0.681	0.538	0.520	0.493	0	0.765	0.514	0.513	0.511	0
	EC ³ -uniM ³	0.688	0.626	0.528	0.503	0	0.769	0.519	0.516	0.515	0
	EC ³ -CQE-uniM ³	0.683	0.561	0.527	0.504	0	0.784	0.583	0.581	0.580	0

Table A9: Performance comparison on the PRP task. The best performance on each task is in **bold**.

Model	IND					
	Acc	M-Pre	M-Rec	M-F1	#Failed	
<i>ft</i> -LLaVA-NExT-Interleave	0.786	0.561	0.243	0.340	2	
eCeLLM-L	0.779	0.558	0.106	0.178	0	
eCeLLM-M	0.775	0.515	0.075	0.131	0	
<i>ft</i> -FashionCLIP	0.738	0.324	0.146	0.201	0	
Task-specific Model	0.779	0.526	0.226	0.316	0	
CASLIE-L	uniM ³	0.785	0.600	0.146	0.235	0
	EC ³ -uniM ³	0.782	0.556	0.177	0.268	0
	EC ³ -CQE-uniM ³	0.782	0.574	0.137	0.221	0
CASLIE-M	uniM ³	0.784	0.557	0.217	0.312	0
	EC ³ -uniM ³	0.783	0.541	0.261	0.352	0
	EC ³ -CQE-uniM ³	0.794	0.586	0.301	0.398	0
CASLIE-S	uniM ³	0.768	0.467	0.190	0.270	0
	EC ³ -uniM ³	0.761	0.443	0.226	0.299	0
	EC ³ -CQE-uniM ³	0.783	0.545	0.243	0.336	0

Table A10: Performance comparison on the PSI task. The best performance on each task is in **bold**.

Model	IND					
	Acc	M-Pre	M-Rec	M-F1	#Failed	
<i>ft</i> -LLaVA-NExT-Interleave	0.721	0.582	0.463	0.469	2	
eCeLLM-L	0.706	0.452	0.431	0.413	0	
eCeLLM-M	0.719	0.467	0.427	0.427	0	
<i>ft</i> -FashionCLIP	0.605	0.372	0.313	0.319	0	
Task-specific Model	0.702	0.469	0.395	0.400	0	
CASLIE-L	uniM ³	0.700	0.446	0.406	0.417	0
	EC ³ -uniM ³	0.704	0.442	0.402	0.411	0
	EC ³ -CQE-uniM ³	0.706	0.708	0.415	0.446	0
CASLIE-M	uniM ³	0.725	0.577	0.500	0.528	0
	EC ³ -uniM ³	0.722	0.596	0.513	0.542	0
	EC ³ -CQE-uniM ³	0.794	0.586	0.301	0.398	0
CASLIE-S	uniM ³	0.699	0.611	0.419	0.445	0
	EC ³ -uniM ³	0.702	0.549	0.448	0.475	0
	EC ³ -CQE-uniM ³	0.707	0.608	0.447	0.481	0

Table A11: Performance comparison on the MPC task. The best performance on each task is in **bold**.

Model	IND					OOD					
	Acc	M-Rec	M-Pre	M-F1	#Failed	Acc	M-Rec	M-Pre	M-F1	#Failed	
<i>ft</i> -LLaVA-NExT-Interleave	0.818	0.577	0.559	0.561	0	0.564	0.208	0.210	0.206	0	
eCeLLM-L	0.830	0.636	0.597	0.613	0	0.827	0.627	0.571	0.584	0	
eCeLLM-M	0.811	0.617	0.652	0.632	0	0.828	0.624	0.629	0.624	0	
<i>ft</i> -FashionCLIP	0.652	0.33	0.318	0.323	0	0.676	0.394	0.379	0.376	0	
Task-specific Model	0.803	0.484	0.525	0.495	0	0.810	0.563	0.535	0.510	0	
CASLIE-L	uniM ³	0.835	0.646	0.616	0.628	0	0.832	0.618	0.588	0.595	0
	EC ³ -uniM ³	0.824	0.613	0.606	0.607	0	0.841	0.648	0.604	0.606	0
	EC ³ -CQE-uniM ³	0.837	0.669	0.640	0.651	0	0.835	0.634	0.600	0.607	0
CASLIE-M	uniM ³	0.839	0.659	0.610	0.617	0	0.850	0.702	0.650	0.659	0
	EC ³ -uniM ³	0.836	0.659	0.631	0.642	0	0.845	0.658	0.609	0.613	0
	EC ³ -CQE-uniM ³	0.845	0.684	0.644	0.656	0	0.846	0.657	0.613	0.625	0
CASLIE-S	uniM ³	0.821	0.564	0.570	0.565	0	0.840	0.662	0.612	0.614	0
	EC ³ -uniM ³	0.825	0.599	0.592	0.578	0	0.831	0.621	0.582	0.565	0
	EC ³ -CQE-uniM ³	0.827	0.616	0.596	0.601	0	0.846	0.690	0.635	0.647	0

Table A12: Performance comparison on the SA task. The best performance on each task is in **bold**.

Model	IND		OOD		
	HR@1	#Failed	HR@1	#Failed	
<i>ft</i> -LLaVA-NExT-Interleave	0.053	0	0.000	0	
eCeLLM-L	0.188	0	0.304	0	
eCeLLM-M	0.182	0	0.302	0	
<i>ft</i> -FashionCLIP	0.145	0	0.087	0	
Task-specific Model	0.163	0	0.210	0	
CASLIE-L	uniM ³	0.184	0	0.285	0
	EC ³ -uniM ³	0.135	21	0.236	0
	EC ³ -CQE-uniM ³	0.190	0	0.297	0
CASLIE-M	uniM ³	0.218	0	0.312	0
	EC ³ -uniM ³	0.207	0	0.310	0
	EC ³ -CQE-uniM ³	0.223	0	0.330	0
CASLIE-S	uniM ³	0.196	0	0.305	0
	EC ³ -uniM ³	0.196	0	0.280	0
	EC ³ -CQE-uniM ³	0.196	0	0.297	0

Table A13: Performance comparison on the SR task. The best performance on each task is in **bold**.

CASLIE models also achieve superior performance over baseline methods in OOD evaluation, demonstrating strong OOD generalizability. Note that in all tables, #failed indicates the number of failure cases for which we cannot extract meaningful results from the model output. We exclude failure cases when calculating the evaluation metrics.

E.5 Case Studies

Case studies are presented in Figure A2, A3, A4, A5, and A6.

F Model Size and Budget

The model size and budget are reported in Table A14.

Model	GPU Memory	Training Time
CASLIE-L	25B	5.0h
CASLIE-M	15B	4.5h
CASLIE-S	7B	3.5h

Table A14: Model budget and size.


Answerability Prediction	
<p>Instruction: Analyze the question and its supporting document, as well as the potential extra information about the products extracted from the product images, predict if the question is answerable based on the provided information. Output only yes or no.</p> <p>Question: Is battery replaceable?</p> <p>Document: ["The battery life is what you would expect from a smart phone. It lasts me with use from about 9pm - 7am", "returned it because the battery life was less than 8 hours even with running almost no apps and it had freezing problems", "the charger was no good it was heating up and the battery doesn't last long I had a few scratches on it and need it a microchip for pictures not enough internal memory", "The phone was fine but the battery had water damage and then got sent another on with water damage. So it never holds a charge", "The battery dies so fast!! waste of my money. Should have just kept my old phone. big big disappointment. Don't order this." ...]</p> <p>Response: yes</p> <p>Generated caption: The image shows a Sprint HTC EVO Shift 4G smartphone box. The battery is replaceable in this model.</p> <p>Is the caption helpful: yes</p>	<p>Product image:</p> 

Figure A2: Case Study of AP



Product Relation Prediction	
<p>Instruction: Given the title of two products, as well as the potential extra information about the products extracted from the product images, predict the relation of the two products. Answer only from the options.</p> <p>Title of product 1: Alfi AB105 Ceramic Wall Mounted Rectangle Bathroom Sink, 17 X 10 X 5 inches, White.</p> <p>Title of product 2: Sliverylake Wall Mount Rectangle White Porcelain Wall-Hung Corner Ceramic Vessel Sink Small & Chrome Faucet Combo.</p> <p>Product image:</p> <p>Options:</p> <ul style="list-style-type: none"> A: Users who buy product 1 may also buy product 2. B: Users who view product 1 may also view product 2. C: The product 1 is similar with the product 2. <p>Response: B</p> <p>Generated caption for product 1: This is a white ceramic wall-mounted rectangular bathroom sink with a single hole for a faucet. It is a modern and minimalist design, suitable for a contemporary bathroom.</p> <p>Generated caption for product 2: A sleek, modern wall-mounted bathroom sink with a chrome faucet, featuring a rectangular design and a clean, white finish.</p> <p>Is the caption helpful: yes.</p>	<p>Product 1 image:</p>  <p>Product 2 image:</p> 

Figure A3: Case Study of PRP

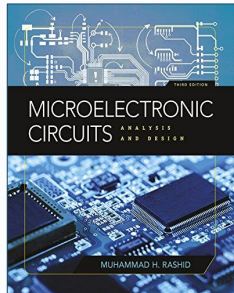
Product Substitute Identification	
<p>Instruction: Given a user's query and a product title, as well as the potential extra information about the product extracted from the product image, identify if the product is somewhat relevant to the query but fails to fulfill some aspects of the query but the product can be used as a functional substitute. Only output yes or no.</p> <p>Query: microelectronics circuit analysis and design 4th edition</p> <p>Product title: Microelectronic Circuits: Analysis and Design</p> <p>Response: yes</p> <p>Generated caption: The third edition of Microelectronic Circuits: Analysis and Design is available.</p> <p>Is the caption helpful: yes</p>	<p>Product image:</p> 

Figure A4: Case Study of PSI


Multiclass Product Classification	
<p>Instruction: Predict the relevance between the query and product by analyzing the user's query, and product title, as well as the potential extra information about the product extracted from the product image. Output the option that best describes the relevance.</p> <p>Query: tree topper angel with moving wings</p> <p>Product title: Conversation Concepts Labrador Retriever Yellow Angel Tree Topper</p> <p>Product image:</p>	
	<p>Options:</p> <p>A: The product is relevant to the query and satisfies all the query specifications.</p> <p>B: The product is somewhat relevant. It fails to fulfill some aspects of the query, but the product can be used as a functional substitute.</p> <p>C: The product does not fulfill the query but could be used in combination with a product exactly matching the query.</p> <p>D: The product is irrelevant to the query.</p> <p>Response: A</p> <p>Generated caption: A Labrador Retriever dressed as a yellow angel with moving wings, designed as a tree topper.</p> <p>Is the caption helpful: yes</p>

Figure A5: Case Study of MPC


Sentiment Analysis	
<p>Instruction: Given the user's review, as well as the potential extra information about the products extracted from the user review's image, identify the user's sentiment. Answer only from the options.</p> <p>Review: Nice handles terrible customer service Impossible to get a hold of the manufacturer by email or phone, I love the product but their is absolutely No way to communicate.</p> <p>Product image:</p>	
	<p>Options:</p> <p>A: Very positive</p> <p>B: Positive</p> <p>C: Neutral</p> <p>D: Negative</p> <p>E: Very negative</p> <p>Response: B</p> <p>Generated caption: A modern kitchen with white cabinets and stainless steel appliances, featuring a large island with a sink and a range hood. The handles on the cabinets are nice, but the customer service from the manufacturer is terrible, making it impossible to get in touch with them.</p> <p>Is the caption helpful: yes.</p>

Figure A6: Case Study of SA