ImplicitAVE: An Open-Source Dataset and Multimodal LLMs Benchmark for Implicit Attribute Value Extraction

Henry Peng Zou[1], **Vinay Samuel**[2], **Yue Zhou**[1], **Weizhi Zhang**[1], **Liancheng Fang**[1], **Zihe Song**[1], **Philip S. Yu**[1], **Cornelia Caragea**[1]
[1]University of Illinois Chicago [2]Carnegie Mellon University
{pzou3,yzhou232,wzhan42,lfang87,zsong29,psyu,cornelia}@uic.edu
vsamuel@andrew.cmu.edu

Abstract

Existing datasets for attribute value extraction (AVE) predominantly focus on explicit attribute values while neglecting the implicit ones, lack product images, are often not publicly available, and lack an in-depth human inspection across diverse domains. To address these limitations, we present ImplicitAVE, the first, publicly available multimodal dataset for implicit attribute value extraction. ImplicitAVE, sourced from the MAVE dataset, is carefully curated and expanded to include implicit AVE and multimodality, resulting in a refined dataset of 68k training and 1.6k testing data across five domains. We also explore the application of multimodal large language models (MLLMs) to implicit AVE, establishing a comprehensive benchmark for MLLMs on the ImplicitAVE dataset. Six recent MLLMs with eleven variants are evaluated across diverse settings, revealing that implicit value extraction remains a challenging task for MLLMs. The contributions of this work include the development and release of ImplicitAVE, and the exploration and benchmarking of various MLLMs for implicit AVE, providing valuable insights and potential future research directions. Dataset and code are available at https://github.com/HenryPengZou/ImplicitAVE.

1 Introduction

Attribute Value Extraction (AVE) identifies the value of product attributes from the product information, which is critical in e-commerce for product representation, recommendation, and categorization (Yang et al., 2022; Wang et al., 2020; Khandelwal et al., 2023; Yang et al., 2023; Fang et al., 2024). The attribute values can be categorized into two types: (1) *Explicit* values can be directly found as a segment in the product text (Yang et al., 2022; Wang et al., 2020), while (2) *Implicit* values are never mentioned in the text and can only be inferred from the product image, contextual clues, or prior knowledge (Zhang et al., 2023). Consider the example in Figure 1. The `value` "rain boot" of the `attribute` "boot style" is implicit since it is not explicitly stated in the product text but can be inferred from its image or context from keywords such as "transparent" and "waterproof."

Figure 1: An example of implicit attribute value. The attribute value "Rain Boot" is not mentioned explicitly in the product text, but can be inferred from text context, product image, or prior knowledge.

Nonetheless, existing datasets for attribute value extraction exhibit several key limitations: (1) They predominantly focus on explicit attribute values, neglecting implicit attribute values (Zheng et al., 2018; Wang et al., 2020), which are more challenging and commonly encountered in real-world scenarios; (2) Many datasets lack product images (Yan et al., 2021; Yang et al., 2022), limiting their applicability in multimodal contexts; (3) The limited number of publicly available datasets lack human inspection and cover only a few domains, resulting in inaccurate and restricted benchmarks (Xu et al., 2019; Zhang et al., 2023). Table 1 compares these aspects for various AVE datasets.

To address these issues, we present ImplicitAVE, the first publicly available multimodal dataset for implicit attribute value extraction. We initially sourced product text data from the MAVE dataset (Yang et al., 2022) and then curated the data by eliminating unhelpful attributes and redundant or irrelevant values. Subsequently, we transformed and

Findings of the Association for Computational Linguistics ACL 2024, pages 338–354

Dataset	Implicit Values	Multimodality	Publicly Available	Human Annotation	Multiple Domains	Language
OpenTag (Zheng et al., 2018)	✗	✗	✗	✓	✓	English
AE-110K (Xu et al., 2019)	✗	✗	✓	✗	✓	Chinese
MEPAVE (Zhu et al., 2020)	✗	✓	✓	✓	✓	Chinese
AdaTag (Yan et al., 2021)	✗	✗	✗	✓	✗	English
MAVE (Yang et al., 2022)	✗	✗	✓	✗	✓	English
DESIRE (Zhang et al., 2023)	✓	✓	✗	✓	✗	Chinese
ImplicitAVE (Ours)	✓	✓	✓	✓	✓	English

Table 1: Comparison of existing AVE datasets. While several ***explicit*** AVE datasets exist, ***implicit*** AVE is much more challenging and under-explored. Our work introduces the first open-source dataset that is expressly designed to address the task of implicit AVE. Our dataset is considerably different from DESIRE, as detailed in Appendix A.

expanded the dataset to include implicit attribute value extraction and multimodality and finally validated the test set annotations through two rounds of human inspection. This yields a more refined and quality-improved dataset of 68k training and 1.6k testing data spanning five diverse domains with 25 attributes and corresponding attribute values suitable for implicit attribute value extraction. Detailed statistics of our dataset are shown in Tables 2, 3.

Given the cutting-edge performance of Multimodal Large Language Models (MLLMs) (Li et al., 2023; Liu et al., 2023b,a; Bai et al., 2023; Ye et al., 2023; Luo et al., 2023) and the absence of previous exploration of their application to implicit attribute value extraction, we establish a comprehensive benchmark for MLLMs on our ImplicitAVE dataset. We cover six recent MLLMs with 11 variants and compare them with the fine-tuned previous SOTA method. We evaluate their performance across diverse settings, including full/few-shot and zero-shot scenarios, domain-level and attribute-level performance, and single/multi-modality performance. We find that implicit value extraction remains a challenging task for open-source MLLMs despite their effective capabilities.

Our contributions are summarized as follows: (1) The development and release of ImplicitAVE, the first open-source multimodal dataset for implicit AVE; (2) The exploration and benchmarking of various MLLMs for implicit attribute value extraction across diverse settings, revealing intriguing insights and potential future research directions.

2 Dataset Construction

We outline our approach to constructing the first open-source multimodal implicit attribute value extraction dataset, ImplicitAVE. The dataset construction pipeline is illustrated in Figure 2. It contains four steps: data collection, curation, expansion, and validation. Next, we explain them in detail.

2.1 Initial Data Collection

Initially, we sourced product text information, including titles, categories, and corresponding attribute-value annotations, from the publicly available MAVE dataset (Yang et al., 2022), comprising 2.2 million products spanning diverse e-commerce domains. Despite its extensive coverage, the MAVE dataset exhibits several significant *limitations*, making it unsuitable for implicit AVE: (1) It contains inappropriate attributes and values that are not facilitative to implicit AVE tasks (see Step 2); (2) It is designed solely for explicit attribute-value extraction; (3) It solely comprises textual information and lacks multimodal data sources; (4) Annotations within the MAVE dataset are machine-generated and lack human inspection, resulting in notable inaccuracies.

2.2 Data Curation for Implicit AVE

We further refine the sourced data by removing unhelpful attributes and redundant or irrelevant values for Implicit AVE. Concretely: ❶ **Removing Inference-Infeasible Attributes**. We manually inspect and remove attributes where the specific values are almost impossible to infer if the values are not mentioned explicitly in the text, such as display resolution, storage capacity, and battery life; ❷ **Removing Subjective Attributes**. The attributes that are rather subjective and ambiguous, such as the degree of comfort and product quality, are also removed; ❸ **Value Merging and Cleaning**. Attribute values with similar semantic meanings are consolidated. This includes unifying variations in grammar forms (e.g., Short-Sleeve, Short sleeves, short sleeved for the attribute Sleeve Style), eliminating extraneous words (e.g., running and running shoes), and merging synonyms (e.g., floral and flower, leopard and cheetah, crew neck and round neck, plaid and tartan, etc.) In addition, we notice some values are irrelevant to their parent

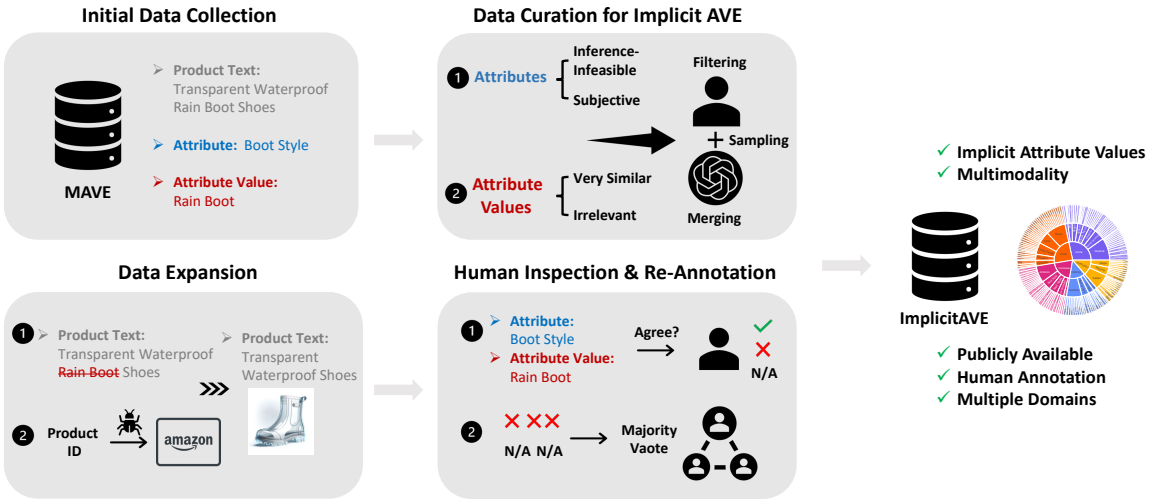


Figure 2: Steps for constructing our ImplicitAVE dataset. A detailed explanation is provided in Section 2.

Domain	# Train	# Eval	# Values	# Attributes	Attributes
Clothing	18868	226	23	4	['Sleeve Style', 'Neckline', 'Length', 'Shoulder Style']
Footwear	21442	317	29	5	['Shaft Height', 'Athletic Shoe Style', 'Boot Style', 'Heel Height', 'Toe Style']
Jewelry&GA	13061	220	20	3	['Pattern', 'Material', 'Shape']
Food	3617	390	41	5	['Form', 'Candy Variety', 'Container', 'Occasion', 'Flavor']
Home Product	11616	457	45	8	['Season', 'Material', 'Location', 'Animal Theme', 'Special Occasion', 'Size', 'Attachment Method', 'Shape']
All	68604	1610	158	25	-

Table 2: Domain-level dataset statistics.

attributes (e.g., the value "Clear Stamps" of the attribute "Material of Artwork"), so these values are removed as well. The value merging and cleaning is achieved collaboratively by lexicon-based scripts, prompting with GPT-4, and human inspection.

This curation results in a more refined and quality-improved dataset with 25 attributes and corresponding attribute values spanning five domains suitable for implicit attribute value extraction. We randomly sample up to 1000 instances per attribute value to limit the dataset size. The selected domains and attributes in ImplicitAVE are shown in Table 2.

2.3 Data Expansion

To extend the data for implicit attribute value extraction and multimodality, we perform the following processing steps: ❶ **Implicit Value Creation**. We remove all explicit attribute value mentions from the input text for its corresponding attribute for each data point. As a result, attribute values in these data can only be inferred from the product images, indirect text context, or prior knowledge. That is, these values become implicit attribute values given the modified inputs. We then drop instances with the same product ID or image to prevent potential information leakage across instances based on the same product. ❷ **Multimodality Creation**. We systematically collect product images from the Amazon website using the product identification number and thus expand our dataset with multimodal information.

2.4 Human Inspection & Re-Annotation

Through manual inspection, we observed that the original attribute-value annotations from MAVE contain noticeable errors. This is because they were annotated by ensembling predictions from five variations of AVEQA models (Wang et al., 2020)[1] without human inspection. To rectify incorrect annotations and ensure a high-quality test set for implicit attribute value extraction and MLLMs evaluation, we engage five Ph.D. students to manually inspect and re-annotate our evaluation set.

This process first involves sampling ten instances per attribute value from the constructed dataset, resulting in 1,676 instances. The human inspection and re-annotation process then unfold in *two* rounds: In the first round, annotators assess each instance's product image, text contexts, and relevant attributes to determine the correctness of the

[1]AVEQA (Wang et al., 2020) is a question-answering model that regards each query attribute as a question and determines the answer span that matches the attribute value within the product text information.

Domains	Attributes	# Train	# Eval	# Values	Attribute Values
Clothing	Sleeve Style	3957	50	5	['Short Sleeve', 'Long Sleeve', '3/4 Sleeve', 'Sleeveless', 'Strappy']
	Neckline	8141	110	11	['Crew Neck', 'V-Neck', 'Henley', 'Polo', 'Scoop Neck', 'Strapless', 'Button Down', ...]
	Length	4937	40	4	['Mini/Short', 'Midi', 'Long Dress/Gown', 'Capri']
	Shoulder Style	1833	26	3	['One Shoulder', 'Off Shoulder', 'Cold Shoulder']
Footwear	Shaft Height	4546	60	5	['Ankle Boot', 'Bootie', 'Knee High', 'Mid Calf', 'Over The Knee']
	Athletic Shoe Style	8165	119	12	['Hiking Boot', 'Soccer', 'Golf', 'Running Shoe', 'Basketball', 'Tennis', 'Walking', ...]
	Boot Style	5145	68	6	['Western/Cowboy', 'Chelsea', 'Combat', 'Snow Boots', 'Motorcycle', 'Rain Boots']
	Heel Height	2457	50	4	['High Heel', 'Flat', 'Mid Heel', 'Low Heel']
	Toe Style	1129	20	2	['Round Toe', 'Pointed Toe']
Jewelry&GA	Pattern	8418	111	10	['Floral', 'Camouflage', 'Plaid', 'Leopard', 'Stripe', 'Paisley', 'Polka Dot', 'Argyle', ...]
	Material	2390	59	5	['Leather', 'Canvas', 'Synthetic', 'Wooden', 'Metal']
	Shape	2253	50	5	['Heart', 'Cross', 'Round', 'Oval', 'Crucifix']
Food	Form	1423	86	9	['Bags/Packets', 'Powder', 'Teabags', 'Rub', 'Bottles', 'Soup Mix', 'Flakes', 'Sticks', 'Sliced']
	Candy Variety	798	82	9	['Gummy/Chewy', 'Gum', 'Hard Candy', 'Mints', 'Licorice', 'Jelly Beans', 'Mint', 'Lollipop']
	Container	563	40	4	['Bag', 'Box', 'Tin', 'Case']
	Occasion	148	43	5	['Easter', 'Other Holiday', "Valentine's", 'Halloween', 'Christmas']
	Flavor	685	139	14	['Vanilla', 'Salted', 'Butter', 'Chocolate', 'Original', 'Strawberry', 'Habanero', 'Caramel', ...]
Home	Season	215	40	5	['All Seasons', 'Autumn', 'Spring', 'Summer', 'Winter']
	Material	7523	158	13	['Metal', 'Ceramic/Melamine', 'Fabric', 'Bamboo', 'Silicone', 'Wood', 'Plastic', 'Glass', ...]
	Location	50	47	4	['Bedroom', 'Kitchen', 'Outdoor', 'Bathroom']
	Animal Theme	134	46	5	['Cat', 'Dog', 'Owl', 'Bird']
	Special Occasion	1002	76	8	['Christmas', 'Halloween', 'Wedding', 'Birthday', 'Graduation', 'Patriotic', 'Easter', ...]
	Size	655	30	4	['Queen', 'King', 'Full', 'Twin']
	Attachment Method	441	20	2	['Grommet', 'Rod Pocket']
	Shape	1596	40	4	['Square', 'Rectangular', 'Oval', 'Round']
All	-	68604	1610	158	-

Table 3: Attribute-level dataset statistics. The detailed ontology of our data and examples of products in different domains, with different attributes and values are provided in Appendix B.

original attribute value annotation. If annotators think the original annotation is incorrect, they select the best attribute value from the corresponding value list (of that attribute) or mark 'N/A' if the annotator believes no suitable value is provided or multiple values are suitable. Additionally, annotators can suggest improvements such as merging, removing, adding, or replacing attribute values. Of the total instances, 1,448 original annotations are correct, 172 are incorrect, and 56 are marked as 'N/A,' yielding an agreement rate of 86.4%. Ten, one, one, and one attribute values are suggested for merging, removing, adding, and replacing, respectively. Instances with disagreed annotations are subject to a second round of inspection and re-annotation, wherein three well-trained annotators participate, and a majority vote determines the final annotation for each instance.

2.5 Dataset Statistics

The overall domain-level dataset statistics is provided in Table 2. We have 68,604 training instances and 1,610 high-quality evaluation instances. Our dataset covers 5 diverse domains and 25 carefully curated attributes specially for the task of implicit attribute value extraction. We also provide detailed attribute-level statistics in Table 3. Different attributes contain different numbers of value options that are meticulously selected and processed and we have a total of 158 diverse attribute values. In addition, we visualize the data distribution of domains, attributes and their values for our training set and evaluation set in Figure 3(a) and 3(b), respectively. It can be observed that compared to the training set, each attribute in the evaluation set has a considerably balanced value distribution, making it more suitable for zero-shot MLLMs evaluation.

3 Experiment & Benchmark

In this section, we describe our experiment results evaluating the effectiveness of various MLLMs and the previous SOTA method on our ImplicitAVE dataset in diverse settings.

3.1 Experimental Setting

Evaluation Setup We benchmark different models on our datasets from both attribute and domain levels:

• **Attribute-Level Results** refer to the micro-F1 score calculated between the ground truth answer and the model-generated answer for *each* query/interested *attribute*.

• **Domain-Level Results** refer to the micro-F1 score calculated between the ground truth answer and the model-generated answer for *all* query/interested *attributes* in *each domain*.

We determine whether the generated answer is correct by checking whether the generated answer contains the true answer.

Method	Language Model	Clothing	Footwear	Jewelry&GA	Food	Home Product	All
Zero-shot methods							
BLIP-2	FlanT5XL-3B	38.05	49.21	72.72	61.54	70.02	59.75
BLIP-2	FlanT5XXL-11B	55.31	55.21	82.72	71.02	71.33	67.39
InstructBLIP	Vicuna-7B	47.79	48.26	76.81	61.28	63.02	59.43
InstructBLIP	FlanT5XXL-11B	**62.83**	63.41	83.18	73.58	73.96	71.49
LLaVA	Vicuna-7B	22.12	39.74	62.72	49.23	57.76	47.82
LLaVA-1.5	Vicuna-7B	26.54	**67.72**	41.95	73.85	66.96	59.69
LLaVA-1.5	Vicuna-13B	49.12	63.72	81.81	76.15	**80.31**	**71.86**
Qwen-VL-Chat	Qwen-7B	32.30	41.01	67.27	55.64	57.11	51.49
Qwen-VL	Qwen-7B	59.73	57.72	**84.09**	**76.92**	73.96	70.86
GPT-4V	-	**77.43**	**81.39**	**90.45**	**90.77**	**89.93**	**86.77**
Representative & SOTA methods (Fine-Tuned)							
DEFLATE (ACL 2023)	T5-Base-770M	54.42	71.61	67.73	52.56	61.71	61.24
LAVIN (NeurIPS 2023)	LLaMA-7B	**65.93**	**75.39**	**78.64**	**60.77**	**64.33**	**67.83**

Table 4: Domain-level results. Analysis and representative error cases are provided in Section 3.2.1. **Bold black** shows best results in each block (zero-shot or finetuning), **bold blue** shows best results overall.

Models for Zero-Shot We utilize the following multimodal LLM frameworks in zero-shot settings:

• **BLIP-2** (Li et al., 2023) proposes a Query Transformer and employs an efficient two-stage vision-and-language pre-training strategy leveraging a frozen image encoder and an LLM. We provide benchmarks of BLIP-2 with two backbone LLM models, FLAN-T5-XL and FLAN-T5-XXL.

• **InstructBLIP** (Dai et al., 2023) enhances vision-language models through instruction tuning with an instruction-aware Query Transformer introduced. We also report the performance with two backbone LLMs, Vicuna-7B and FLAN-T5-XXL.

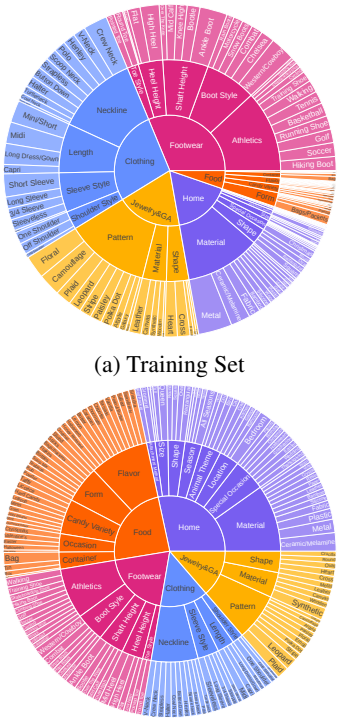
• **LLaVA** (Liu et al., 2023b) connects the visual encoder of CLIP (Radford et al., 2021) with the language decoder, and performs fine-tuning on GPT-4 generated language-image instructions. We provide benchmarks of LLaVA with Vicuna-13B.

• **LLaVA-1.5** (Liu et al., 2023a) advances its predecessor by focusing on efficient visual instruction tuning, integrating a fully-connected vision-language cross-modal connector for enhanced interaction between visual and textual modality. We provide benchmarks of LLaVA-1.5 using Vicuna-7B and Vicuna-13B as the language models.

• **Qwen-VL** (Bai et al., 2023) proposes a novel visual receptor and a position-aware adapter, optimizing through a three-stage training pipeline on a multilingual and multimodal dataset. We report the performance of both Qwen-VL and the chat version, Qwen-VL-Chat.

• **GPT-4V**[2] integrates vision into the GPT-4 architecture, one of the cutting-edge close-sourced LLMs fine-tuned by reinforcement learning from human feedback.

[2]https://chat.openai.com/



(a) Training Set

(b) Evaluation Set

Figure 3: Data distribution of domains, attributes, and attribute values for training and evaluation sets. (A full-size version is attached to our appendix - Figure 10)

Models for Finetuning Due to the resource constraints, we fine-tuned and evaluated the following two open-source models in both few-shot and full-data tuning settings:

• **LaVIN** (Luo et al., 2023) introduces a novel mix-of-modularity adaptation module, optimizing the integration of visual inputs into large language models through lightweight adapters and enabling efficient end-to-end training.

• **DEFLATE** (Zhang et al., 2023) is a multi-modal

Domains	**Attributes**	**# Values**	InstructBLIP	LLaVA 1.5	Qwen-VL	GPT-4V	DEFLATE	LAVIN
Language Model/Variants			*FlanT5XXL-11B*	*Vicuna 13B*	*Qwen-7B*	-	*T5-Base-770M*	*LLaMA-7B*
Food	Flavor	14	72.66	84.17	89.21	**97.12**	51.08	53.24
Home	Material	13	74.05	61.39	67.09	**84.81**	77.22	82.28
Jewelry&GA	Pattern	10	81.08	80.18	89.19	**90.99**	61.26	78.38
Footwear	Athletic Shoe Style	12	73.95	63.03	57.98	**84.03**	80.67	78.15
Clothing	Neckline	11	53.64	25.45	52.73	**78.18**	50.91	57.27
Food	Form	9	70.93	59.30	75.58	**86.05**	67.44	81.40
Home	Special Occasion	8	90.79	92.11	88.15	**98.68**	72.37	68.42
Clothing	Sleeve Style	5	62.00	46.00	66.00	66.00	34.00	**70.00**
Footwear	Boot Style	6	76.47	73.53	72.05	**88.24**	75.00	83.82
Jewelry&GA	Material	5	81.36	93.22	88.14	**94.92**	77.97	86.44
Food	Container	4	87.50	**95.00**	80.00	87.50	52.50	60.00
Footwear	Heel Height	4	58.00	54.00	54.00	**86.00**	62.00	72.00
Clothing	Shoulder Style	3	**88.46**	42.31	80.77	80.77	69.23	61.54
Home	Attachment Method	2	45.00	100.00	100.00	**100.00**	90.00	90.00

Table 5: Attribute-level results. Analysis and representative error cases are provided in Section 3.2.2. Best results per attribute are shown in **bold blue**.

generative-discriminative framework designed for both explicit and implicit attribute value extraction and is the previous SOTA model for implicit AVE.

3.2 Experimental Results

3.2.1 Domain-Level Results

We present the domain-level results of all evaluated models in Table 4. GPT-4V outperformed every other model in both the zero-shot and finetune setting in every single domain. Among the two models that were finetuned, LAVIN outperformed DEFLATE in every single domain by a minimum of 2.62 points (in the Home Product domain) and a maximum of 11.51 (in the Clothing domain). Among the open-source MLLMs, no single model outperformed all other models across all the domains, but Qwen-VL had the best scores in the Jewelry&GA and Food domains. From Table 4 we also note that other than for LLaVA 1.5 in the Footwear domain, all other models that had multiple variants with different LLM sizes had significantly better performance on average from the variant with the larger size LLM in each domain in comparison to the variant with the smaller sized LLM. For example, in the Clothing domain, there was a minimum improvement of 15.04 micro-F1 points from the model variant with the smaller LLM (InstructBLIP w/ Vicuna 7B) to the model variant with the larger LLM (InstructBLIP w/ FLAN-T5-XXL) and an overall average of 18.29 micro-F1 point increase when using a model variant with a larger LLM in the Clothing domain. Similar trends can be seen among all domains.

Additionally, among zero-shot methods, Clothing had the lowest micro-F1 across all domains for all models and model variants except for BLIP2 w/ FLAN-T5-XL and Qwen-VL. This leads us to believe that the Clothing domain is the most challenging domain in the dataset. We performed a comprehensive manual investigation and we believe there are two *primary reasons why the Clothing domain presents more challenges*, while other domains such as the Home domain are comparatively easier (We show examples from our manual investigation in Figures 8, 9 for clarity):

(1) Attributes within the Clothing domain demand a more nuanced understanding of local details in product images. For example, the attribute 'Sleeve Style' in cases 1-4 and 'Neckline' in cases 7-12 (Figure 8). In contrast, attributes in the home domain only require a global understanding of product pictures and text, such as attribute 'Special Occasion' in cases 13-16, 'Shape' and 'Material' in cases 17 and 21 (Figure 9).

(2) The values of attributes in the Home domain are significantly more straightforward to identify compared to those in the Clothing domain. For instance, the attribute 'Special Occasion' includes values like ['Birthday', 'Christmas', 'Easter', 'Graduation', 'Halloween', 'Patriotic', 'Thanksgiving'], which are clearly more distinguishable than the values for 'Sleeve Style' ['Sleeveless', 'Long Sleeve', '3/4 Sleeve', 'Strappy', 'Short Sleeve'] in the Clothing Domain.

3.2.2 Attribute-Level Results

Table 5 presents the attribute-level performance of all evaluated models. As was observed in Table 4, GPT-4V vastly outperforms all other models. We can see in Table 5 that only in the 'Shoul-

Domains	GPT-4V	Qwen-VL	LLaVA-1.5	InstructBLIP	BLIP-2
Clothing	77.43	59.73	49.12	62.83	55.31
Footwear	81.39	57.72	67.72	63.41	55.21
Attributes					
Sleeve Style	66.00	66.00	46.00	62.00	50.00
Shaft Height	63.33	35.00	61.66	26.67	30.00
Season	65.00	57.50	65.00	60.00	62.50
Neckline	78.18	52.73	25.45	53.64	48.18
Average	**68.13**	52.81	49.53	50.58	47.67

Table 6: Examples of challenging domains & attributes.

der Style' (InstructBLIP), 'Container' (LLaVA 1.5) and 'Sleeve Style' (LAVIN) attributes do a model outperform GPT-4V. InstructBLIP struggled significantly with the 'Attachment Method' attribute as did LLaVA 1.5 with 'Shoulder Style' compared to other models. On the other hand, Table 5 shows that both finetuned models perform better than all of the open-source MLLMs in the zero-shot setting for the 'Heel Height' attribute. This may indicate that there are attributes within the dataset for which prior pretrained knowledge of MLLMs is not sufficient for implicit value extraction of these attribute values and finetuning is needed to learn the mapping between instances of these attributes and the correct attribute values belonging to them.

In addition, all models struggled on the 'Sleeve Style' and 'Neckline' attributes compared to each model's performance on other attributes. Representative error cases for different attributes are presented in Figures 8 and 9 in Appendix C along with a comprehensive error analysis. Here we provide our observations from the *attribute-level error analysis*:

(1) Models often confuse attribute values that are similar yet distinct, such as '3/4 Sleeve' versus 'Long Sleeve' in cases 1-2, 'Short Sleeve' versus 'Sleeveless' in cases 3-4, and 'Crew Neck' versus 'Scoop Neck' in case 8 (Figure 8).

(2) Attributes that demand a detailed understanding of small image parts typically challenge models, leading to errors. For instance, mistakes in identifying 'Shoulder Style' in cases 5-6 and 'Neckline' in cases 7-9 (Figure 8).

(3) Errors can also arise from conflicting modality inferences, as seen in case 13 (Figure 9), where the word 'Snow Village' in the product text suggested Christmas, but the image aligned more with Halloween.

3.2.3 Challenges and Opportunities

Challenging Domains & Attributes: It can be observed in Table 4, 5 that GPT-4V works well on some domains and attributes, but not on all of them, e.g., it only achieves 77.4% micro-F1 on the Clothing domain and 66.0% for the Sleeve Style attribute. Some examples of challenging domains, attributes, and the performance of various MLLMs are highlighted in Table 6. Besides, we can observe that the open-source models are lagging behind GPT-4V in many domains and attributes, and our dataset provides a good benchmark that points out the gap between them and provides opportunities for researchers to close it.

Furthermore, inspired by the error cases in Section 3.2.2 and Appendix C, we point out some *remaining challenges and opportunities*:

Model-Aspect: (1) Enhance the ability to understand image details, including small areas and text in images; (2) Devise mechanisms to distinguish similar attribute values; (3) Properly handle conflicting modality inferences; (4) Reduce the performance gap in implicit AVE between open-source models and advanced closed models like GPT-4V.

Dataset-Aspect: Our ImplicitAVE dataset does not consider multi-valued attributes and negative instances, i.e., "none" as attribute values. We leave this extension for future work.

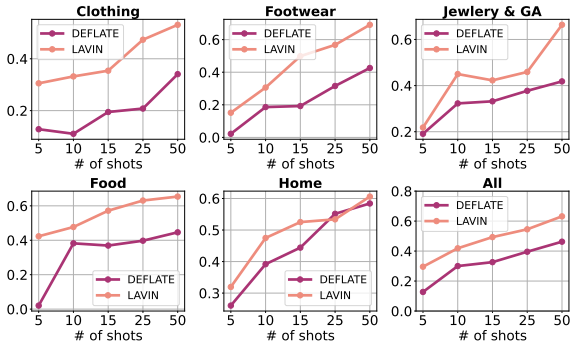
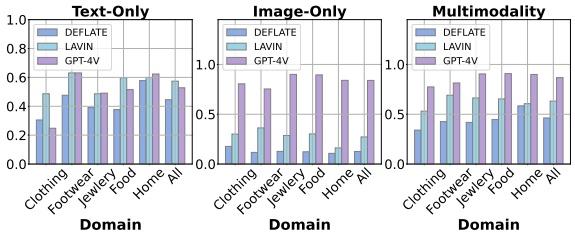


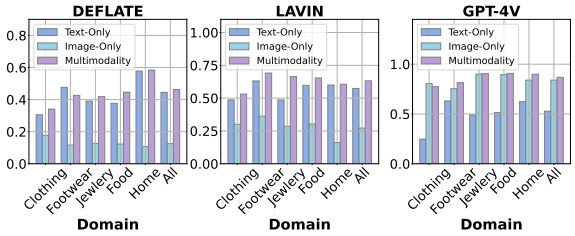
Figure 4: Performance comparison in few-shot settings of different domains.

3.2.4 Few-Shot Results

Figure 4 shows the performance comparison of DEFLATE and LAVIN models in various few-shot settings. We note that in most K-shot settings, LAVIN outperforms DEFLATE by a noticeable amount. Also, we notice that different domains performed differently for the two models. In the 5-shot setting, 'Food' was the lowest scoring do-



(a) Comparison of methods.



(b) Comparison of modality.

Figure 5: Performance comparison of DEFLATE, LAVIN, and GPT-4V on different modalities.

main for DEFLATE whereas 'Food' was the highest scoring domain for LAVIN, but for 10-shot the domain trends for both models became similar (i.e., 'Food' and 'Home Product' domains were the two best-performing domains and 'Clothing' and 'Footwear' were the worst performing domains). For DEFLATE from 25-shot to 50-shot, the largest increase in micro-F1 was for the 'Clothing' and 'Footwear' domains whereas the increase was less significant for the other domains. This indicates that the model's ability to learn the attributes and attribute values in the 'Clothing' and 'Footwear' domains may continue to increase as the number of training examples increases. On the contrary, LAVIN saw the biggest increase in micro-F1 for the 'Jewelry&GA' and 'Footwear' domains thereby hinting that increasing the training examples for these domains in LAVIN would enable the model to substantially increase its ability to categorize instances of these two domains.

3.2.5 Modality-Level Results

Figure 5 visualizes performance comparisons of DEFLATE, LAVIN, and GPT-4V with different modalities. Firstly, it is evident that for LAVIN and DEFLATE, the image-only modality performed extremely poorly compared to the text-only and combined modalities. This leads us to believe that these models' image understanding capabilities may be too poor to extract implicit value from product images. However, it is worth noting that in all domains except 'Footwear' for DEFLATE, both LAVIN and DEFLATE perform better in the multimodal modality over the text-only modality thereby indicating that the image information does in fact help the model predict attribute values of instances. With GPT-4V we notice a very high performance in the image-only modality and only minimal improvement with the multimodal modality in comparison to the image-only modality. This speaks to the strength of GPT-4V's zero-shot image classification capabilities, especially in comparison to LAVIN and DEFLATE. Even though GPT-4V boasts impressive performance in most regards, it is worth noting that GPT-4V's text-only modality performance in the 'Clothing' domain was especially poor. It scored even lower than the text-only scores of LAVIN and DEFLATE and, in the 'Clothing' domain, the multimodal performance for GPT-4V was lower than the image-only performance thereby indicating that the text component confused the model causing it to perform worse than it did without the text component.

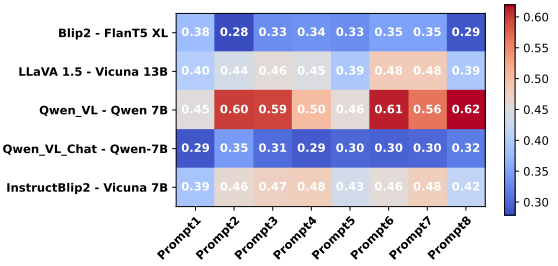


Figure 6: The influence of prompts (detailed in Table 7 in Appendix E) on different models.

3.3 Ablation Study on Prompt Templates

In order to obtain baseline results that accurately reflect the quality of our dataset we conducted ablations on the prompt for the open-source MLLMs. Observing drastic micro-F1 score differences on the evaluation set by using different prompts in the early stages of experimenting led us to conduct a standardized ablation study on 8 different prompts listed in Table 7. Each prompt had three components: context containing the title of the product with explicit mention of the attribute value removed, question, and options to answer from. To conduct a fair evaluation of the prompts, across all models we fixed the random seed at 42 as well as the hyperparameters: temperature = 1, top_p = 0.8, max_new_tokens = 17, min_length = 1, and num_beams = 5. Our results are shown in Fig 6 and the best prompt for each model type was used for all variants of that model.

4 Related Work

4.1 Attribute Value Extraction Dataset

Attribute Value Extraction (AVE) has emerged as a crucial task for online shopping, aiming to identify the values of product attributes from various data sources. At the heart of many e-commerce applications, such as product comparison, retrieval, recommendation, and the construction of product graphs and online shop assistants, lies the extraction of attribute values (Zalmout et al., 2021). Although several AVE datasets have been introduced, each exhibits certain limitations, as shown in Table 1.

The OpenTag dataset (Zheng et al., 2018), one of the early datasets collected from Amazon, highlights the importance of open-world value sets. In contrast, the AE-110K dataset (Xu et al., 2019) expands the scope of AVE datasets to include more products, a broader range of attributes, and denser attribute coverage per product, though it lacks human expert annotation. The AdaTag dataset (Yan et al., 2021) focuses on the rich information contained in product bullets, excluding product descriptions, which facilitates more efficient training and inference for such tasks but lacks diversity in product domains and is not publicly available. The MAVE dataset (Yang et al., 2022), a large public dataset for AVE research, encompasses a wide range of categories and diverse attributes, constructing structured product files as text inputs. However, in real-world scenarios, text information alone may not imply certain attributes of interest, making product images a complementary source of information for indicating or validating the answers to specific attributes. To address this, the MEPAVE (Zhu et al., 2020) and DESIRE (Zhang et al., 2023) datasets were introduced to include multimodal product information such as product titles, descriptions and images. While several explicit AVE datasets exist, implicit AVE is much more challenging and under-explored. To advance multi-modal AVE research further, we introduce the first publicly available multimodal implicit AVE dataset, ImplicitAVE, featuring careful human annotation and a versatile range of items from multiple domains. Our dataset is considerably different from DESIRE, as detailed in Appendix A.

4.2 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) have demonstrated impressive performance on a variety of tasks (Li et al., 2023; Liu et al., 2023b,a; Bai et al., 2023; Ye et al., 2023; Luo et al., 2023; Dong et al., 2024). BLIP-2 (Li et al., 2023) uses frozen pre-trained image models and language models, and proposes a lightweight querying transformer Q-Former to bridge the two modalities. InstructBLIP (Dai et al., 2023) outperforms BLIP-2 (Li et al., 2023) by using vision-language instruction tuning, where the instruction tuning data is collected from publicly available datasets, by manually transforming them into instruction tuning format. To improve the diversity and in-depth reasoning in the instruction, LLaVa (Liu et al., 2023b) proposes to use language-only GPT-4 to construct multimodal language-image instruction tuning data. mPLUG-Owl (Ye et al., 2023) and Qwen-VL (Bai et al., 2023) propose novel training paradigms for LLMs. However, since most popular open-source MLLMs are parameter-heavy, LAVIN (Luo et al., 2023) proposes a novel and efficient solution for vision-language instruction tuning by adopting lightweight modules, i.e., adapters, to bridge the gap between LLMs and vision modules, which does not require expensive vision-language pretraining to align text and image embedding beforehand. Despite achieving significant progress, the performance of MLLMs on implicit AVE has not been well-studied. Recent work EIVEN (Zou et al., 2024) finetuned an efficient MLLM framework for implicit AVE but did not compare with exiting MLLMs in zero-/few-shot settings. Our work establishes the first comprehensive benchmark of multimodal LLMs for implicit AVE under diverse settings and reveals intriguing insights and potential future research directions in Section 3.2.3.

5 Conclusion

In this paper, we introduced ImplicitAVE, the first publicly accessible multimodal dataset specifically designed for implicit attribute value extraction, aimed at overcoming the limitations of existing datasets focused on explicit attribute values. By carefully curating attribute values and incorporating both implicit attribute values and product images, ImplicitAVE comprises 6.8K training instances and 1.6K human re-annotated high-quality evaluation instances across five diverse domains. Moreover, we benchmarked the performance of six recent multimodal large language models on it under diverse settings, highlighting the challenges of implicit value extraction. In the future, we plan to further expand our ImplicitAVE dataset to include multi-valued attributes and negative instances.

Acknowledgements

This research is partially supported by NSF grant #210751 and UIC DPI Seed Program. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF. We thank our reviewers for their insightful feedback and comments which helped improve the quality of our paper.

6 Limitation

Our ImplicitAVE dataset does not consider multi-valued attributes and negative instances, i.e. "none" as attribute values. We leave this extension as future work. Due to computational resource constraints and limited budgets, we did not evaluate open MLLMs with parameters larger than 13B.

7 Ethics Statement

The datasets that we sourced from are publicly available. In this work, we propose a multimodal Implicit AVE dataset and provide a comprehensive benchmark of MLLMs. We do not expect any direct ethical concern from our work.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhikang Dong, Bin Chen, Xiulong Liu, Pawel Polak, and Peng Zhang. 2024. Musechat: A conversational music recommendation system for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024. Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. *arXiv preprint arXiv:2403.00863*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. 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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D. Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 47–55, New York, NY, USA. Association for Computing Machinery.

Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223, Florence, Italy. Association for Computational Linguistics.

Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4694–4705, Online. Association for Computational Linguistics.

Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, Sinong Wang, Zenglin Xu, and Dongfang Liu. 2023. MixPAVE: Mix-prompt tuning for few-shot product attribute value extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9978–9991, Toronto, Canada. Association for Computational Linguistics.

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*, WSDM '22, page 1256–1265, New York, NY, USA. Association for Computing Machinery.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Nasser Zalmout, Chenwei Zhang, Xian Li, Yan Liang, and Xin Luna Dong. 2021. All you need to know to build a product knowledge graph. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 4090–4091, New York, NY, USA. Association for Computing Machinery.

Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023. Pay attention to implicit attribute values: A multi-modal generative framework for AVE task. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13139–13151, Toronto, Canada. Association for Computational Linguistics.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1049–1058, New York, NY, USA. Association for Computing Machinery.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2129–2139, Online. Association for Computational Linguistics.

Henry Peng Zou, Gavin Heqing Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024. Eiven: Efficient implicit attribute value extraction using multimodal llm. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*.

A Novelty and Contribution of Dataset

In Table 1, we compare our dataset with existing AVE datasets from different aspects. While several ***explicit*** AVE datasets exist, ***implicit*** AVE is much more challenging and underexplored. To the best of our knowledge, our work introduces the first open-source dataset that is expressly designed to address the task of implicit AVE. Here, we would like to clarify that our dataset is **considerably different** from DESIRE (Zhang et al., 2023) regarding:

1. **Accessibility:** In DESIRE, all data are encrypted, and image data are encoded by DALL-E (Ramesh et al., 2021), with raw images unavailable. Moreover, DESIRE does not provide statistics on how many implicit AVE examples are included. In contrast, our dataset contains 70k+ curated implicit AVE examples, which are immediately available with raw images provided.

2. **Domain Scope:** DESIRE only contains the 'Food' domain, while our dataset contains five domains, including more challenging domains such as 'Clothing' and 'Footwear'.

3. **Language:** DESIRE is based on Chinese while our dataset is in English.

B Domain, Attribute, and Value

The **ontology** of our data adheres to the ***domain-attribute-value*** structure, where (1) Each domain contains relevant attributes that characterize different aspects of the domain product. For instance, the 'Clothing' domain contains attributes such as 'Sleeve Style' and 'Neckline'; (2) Each attribute comprises a set of possible values (also called "attribute values") and we aim to extract its ground truth value from product images and text contexts. For example, the attribute 'Sleeve Style' may include values such as 'Long-sleeve', '3/4 sleeve', and 'Strappy'. The full details are depicted in Table 3. Figure 7 also presents a few examples of products in different domains, with different attributes and values.

Therefore, in Tables 4 and 5, **Attribute-level results** refer to the micro-F1 score calculated between the ground truth answer and the model-generated answer for each query/interested attribute. **Domain-level results** refer to the micro-F1 score calculated between the ground truth answer and the model-generated answer for all query/interested attributes in each domain. We determine whether the generated answer is correct by checking whether the generated answer contains the true answer.

C Detailed Error Analysis and Remaining Challenges

We have conducted an exhaustive analysis of cases incorrectly predicted by various models, with a particular focus on GPT-4V. **Representative error cases** for different domains and attributes are presented in Figure 8 and Figure 9. Here we provide a more detailed error analysis from different perspectives:

Attribute-Level: (1) Models often confuse attribute values that are similar yet distinct, such as '3/4 Sleeve' versus 'Long Sleeve' in cases 1-2, 'Short Sleeve' versus 'Sleeveless' in cases 3-4, and 'Crew Neck' versus 'Scoop Neck' in case 8. (2) Attributes that demand a detailed understanding of small image parts typically challenge models, leading to errors. For instance, mistakes in identifying 'Shoulder Style' in cases 5-6 and 'Neckline' in cases 7-9. (3) Errors can also arise from conflicting modality inferences, as seen in case 13, where the word 'Snow Village' in the product text suggested Christmas, but the image aligned more with Halloween.

Domain-Level: (1) 'Clothing' is the hardest domain for most models because it contains many attributes that require fine-grained understanding of product images. For example, 'Sleeve Style' in cases 1-4 and 'Neckline' in cases 7-12. (2) While easier domains such as 'Home' usually consist of attributes that only need a more global understanding of product image and text context, such as 'Special Occasion' in cases 13-16, 'Shape' and 'Material' in cases 17 and 21.

Model-Level: (1) Open-source models are not good at recognizing and leveraging text in images. Taking case 19 as an example, Qwen-VL, DEFLATE, and LAVIN fail to utilize the text words 'BATHROOM, Teeth, Toilet' in the image. (2) Interestingly and uniquely, when GPT-4V considers none of the provided options suitable, it will answer 'None' and then give an answer it feels is a better match, as shown in cases 6, 18. (3) LLaVA 1.5 tends to provide multiple answers in ambiguous situations, as can be seen in cases 16, 18.

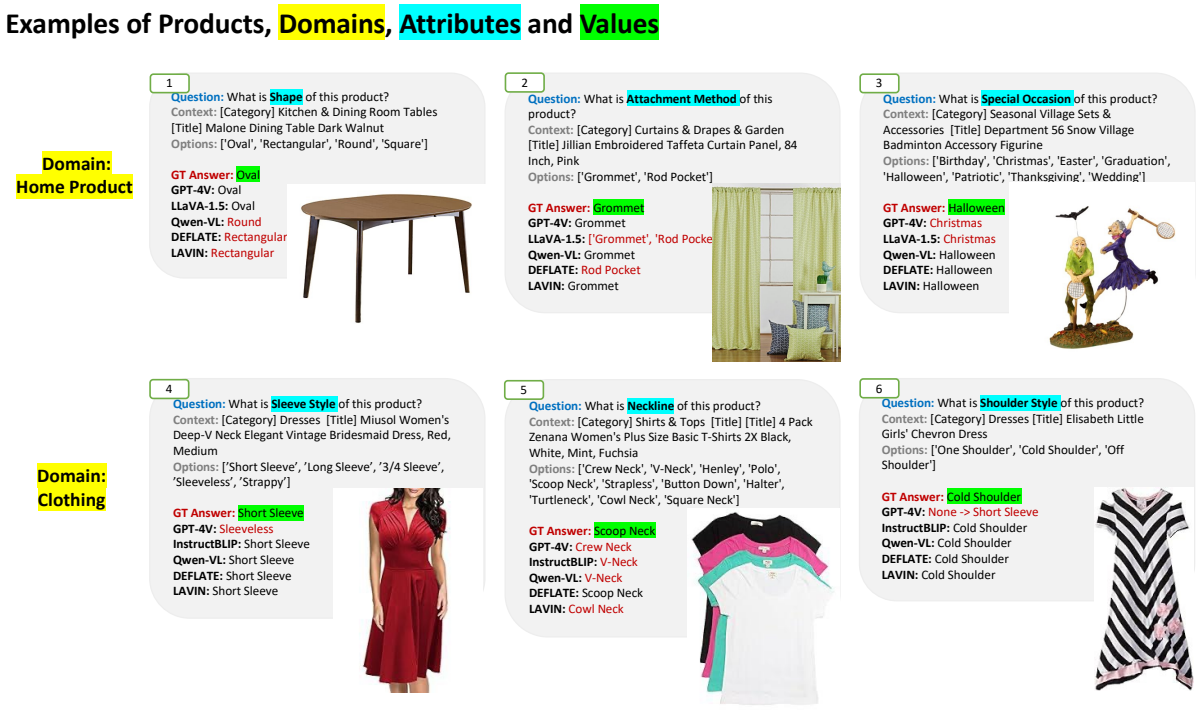


Figure 7: Examples of products, domains, attributes, values.

Inspired by the above error cases, we point out some **remaining challenges and opportunities**:

Model-Aspect: (1) Enhance the ability to understand image details, including small areas and text in images; (2) Devise mechanisms to distinguish similar attribute values; (3) Properly handle conflicting modality inferences; (4) Reduce the performance gap in implicit AVE between open-source models and advanced closed models like GPT-4V.

Dataset-Aspect: Our ImplicitAVE dataset does not consider multi-valued attributes and negative instances, i.e. "none" as attribute values. We leave this extension for future work.

D Implementation Details

All open-source MLLMs are evaluated on a single A100 GPU. Due to RAM and Disk space constraints, all model pre-trained weights were loaded in at 4 or 8 bits using the bitsandbytes library for quantization. All model variants using the Vicuna-13B or Flan-T5-XXL are loaded in at 4 bits, and all other models are loaded in at 8 bits. Additionally, eight different prompts are tested, with the best performance reported for each open-source model (BLIP2, InstructBLIP, LLaVA, LLaVA 1.5, Qwen-VL). A list of valid attribute value options is provided when prompting the MLLMs. For all settings, we use micro-F1/accuracy as evaluation metrics. The prompt templates are available in the Appendix E.

E Prompt Templates

Table 7 provides our prompt templates for all zero-shot methods except GPT-4V. The best results are displayed. The prompt template we use for GPT-4V is: "What is the {attribute_names} of this product? Context: [Category] {category} {texts}. Choose the most appropriate one from the options: {Options}."

Prompt 1	"Question: What is {attribute_names} of this product?\nContext: [Category] {category} {texts}.\n You must only answer the question with exactly one of the following options {options}.\nAnswer:"
Prompt 2	"What is {attribute_names} of this product?[Category] {category} {texts}. Answer with the option from the given choices directly: {options}.\nAnswer:"
Prompt 3	"[Category] {category} {texts}. What is {attribute_names} of this product? Answer with the option from the given choices directly: {options}.\nAnswer:"
Prompt 4	"[Category] {category} {texts}. What is {attribute_names} of this product based on the given information and the given image? Answer with the option from the given choices directly: {options}.\nAnswer:"
Prompt 5	"[Category] {category} {texts}. Which one of {options} is the {attribute_names} of this product? Answer with the option from the given choices directly.\nAnswer:"
Prompt 6	"{texts}. What is the {attribute_names} of this product? Answer with the option from the given choices directly: {options}.\nAnswer:"
Prompt 7	"{texts}. Based on the description and the image, what is the {attribute_names} of this product? Answer with the option from the given choices directly: {options}.\nAnswer:"
Prompt 8	"What is the {attribute_names} of this product: {texts}? Answer with the option from the given choices directly: {options}.\nAnswer:"

Table 7: Prompt templates for all zero-shot methods except GPT-4V. The best prompt for each model type was used for all variants of that model. The prompt used for GPT-4V is provided in Appendix E and our code.

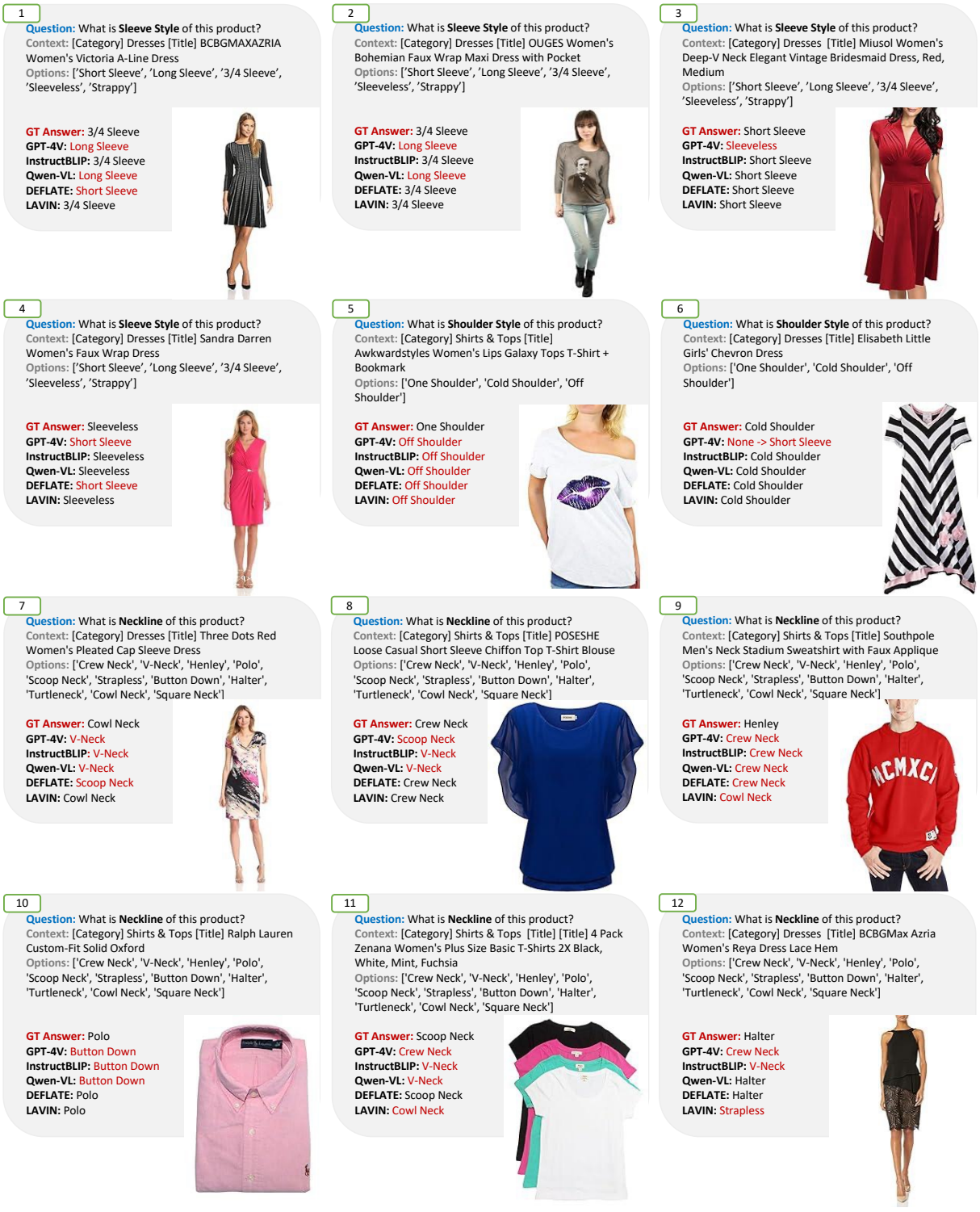


Figure 8: Representative error cases - clothing domain. (Domain-level analysis: Section 3.2.1; Attribute-level analysis: Section 3.2.2; Comprehensive error analysis and remaining challenges: Appendix C)

13

Question: What is **Special Occasion** of this product?
Context: [Category] Seasonal Village Sets & Accessories [Title] Department 56 Snow Village Badminton Accessory Figurine
Options: ['Birthday', 'Christmas', 'Easter', 'Graduation', 'Halloween', 'Patriotic', 'Thanksgiving', 'Wedding']

GT Answer: Halloween
GPT-4V: Christmas
LLaVA-1.5: Christmas
Qwen-VL: Halloween
DEFLATE: Halloween
LAVIN: Halloween

14

Question: What is **Special Occasion** of this product?
Context: [Category] Artwork [Title] Snowbunnies Unraveling Figurine
Options: ['Birthday', 'Christmas', 'Easter', 'Graduation', 'Halloween', 'Patriotic', 'Thanksgiving', 'Wedding']

GT Answer: Easter
GPT-4V: Easter
LLaVA-1.5: Easter
Qwen-VL: Christmas
DEFLATE: Christmas
LAVIN: Easter

15

Question: What is **Special Occasion** of this product?
Context: [Category] Paper Napkins [Title] C.R. Gibson 20 Count Autumn Array Dinner Napkins
Options: ['Birthday', 'Christmas', 'Easter', 'Graduation', 'Halloween', 'Patriotic', 'Thanksgiving', 'Wedding']

GT Answer: Thanksgiving
GPT-4V: Thanksgiving
LLaVA-1.5: Halloween
Qwen-VL: Thanksgiving
DEFLATE: Christmas
LAVIN: Thanks

16

Question: What is **Special Occasion** of this product?
Context: [Category] Vases [Title] Lenox Promises Opal Innocence Bud Vase
Options: ['Birthday', 'Christmas', 'Easter', 'Graduation', 'Halloween', 'Patriotic', 'Thanksgiving', 'Wedding']

GT Answer: Wedding
GPT-4V: Wedding
LLaVA-1.5: ['Birthday', 'Christmas', 'Easter', 'Graduation']
Qwen-VL: Wedding
DEFLATE: Christmas
LAVIN: Wedding

17

Question: What is **Shape** of this product?
Context: [Category] Kitchen & Dining Room Tables [Title] Malone Dining Table Dark Walnut
Options: ['Oval', 'Rectangular', 'Round', 'Square']

GT Answer: Oval
GPT-4V: Oval
LLaVA-1.5: Oval
Qwen-VL: Round
DEFLATE: Rectangular
LAVIN: Rectangular

18

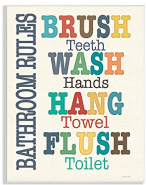
Question: What is **Shape** of this product?
Context: [Category] Mirrors [Title] Abbyson Pulse Shaped Wall Mirror
Options: ['Oval', 'Rectangular', 'Round', 'Square']

GT Answer: Square
GPT-4V: None -> Heart
LLaVA-1.5: ['Oval', 'Rectar 'Round', 'Square']
Qwen-VL: Round
DEFLATE: Oval
LAVIN: Square

19

Question: What is **Location** of this product?
Context: [Category] Home & Garden [Title] The Stupell Home Decor Collection Colorful Rules Typog Wall Plaque Art, 10 x 0.5 x 15
Options: ['Bathroom', 'Bedroom', 'Garage', 'Kitchen']

GT Answer: Bathroom
GPT-4V: Bathroom
LLaVA-1.5: Bathroom
Qwen-VL: Kitchen
DEFLATE: Kitchen
LAVIN: Kitchen



20

Question: What is **Attachment Method** of this product?
Context: [Category] Curtains & Drapes & Garden [Title] Jillian Embroidered Taffeta Curtain Panel, 84 Inch, Pink
Options: ['Grommet', 'Rod Pocket']

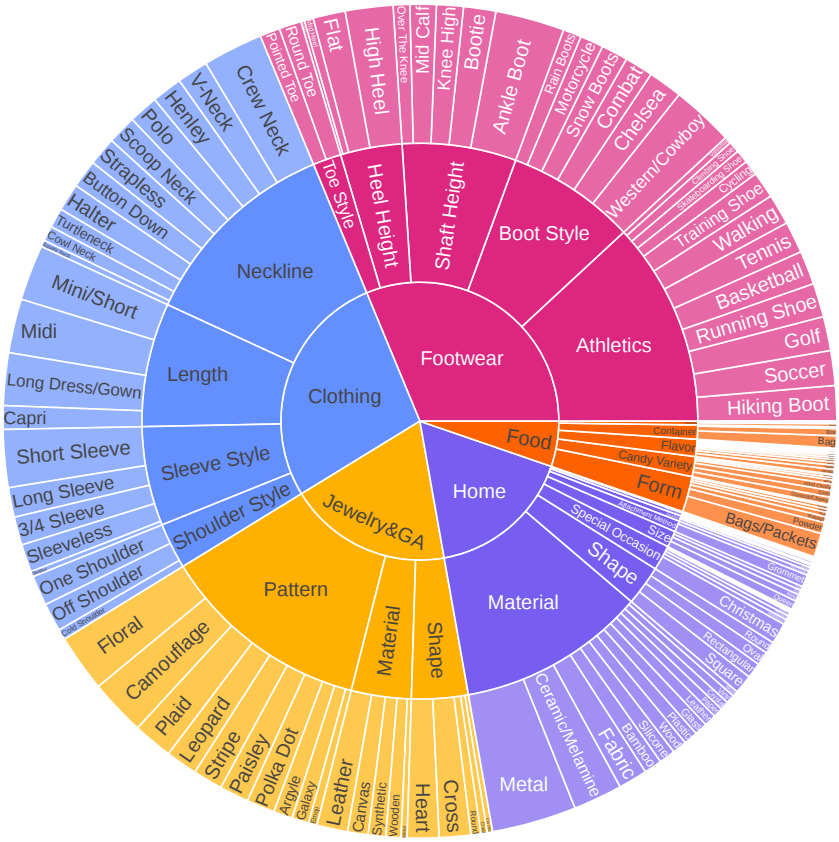
GT Answer: Grommet
GPT-4V: Grommet
LLaVA-1.5: ['Grommet', 'Rod Pocke
Qwen-VL: Grommet
DEFLATE: Rod Pocket
LAVIN: Grommet

21

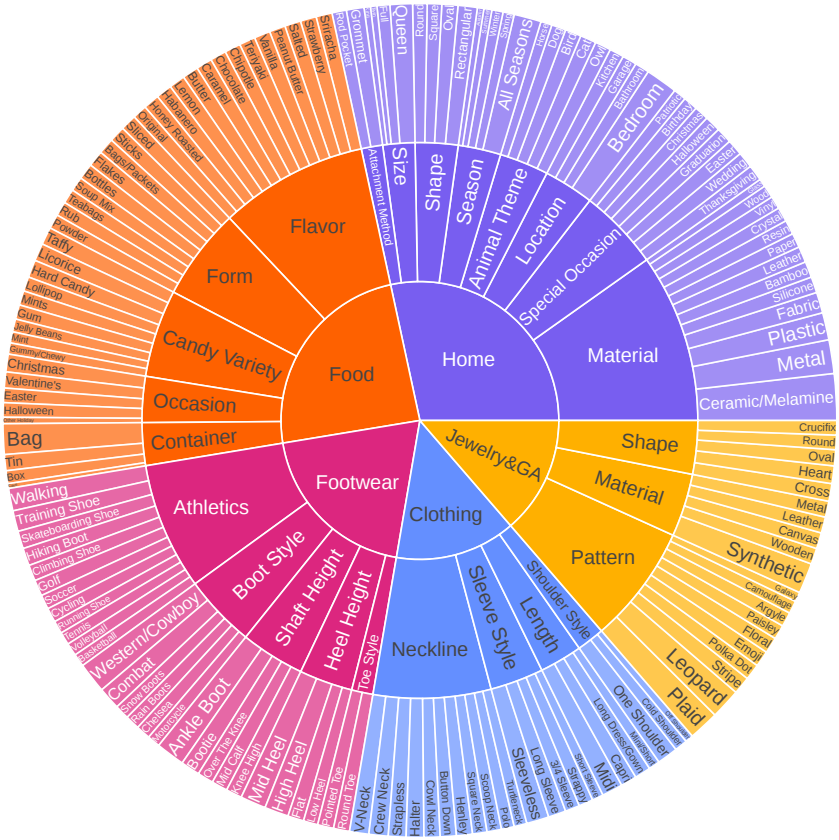
Question: What is **Material** of this product?
Context: [Category] Food Storage Containers [Title] Oggi Airtight 3 Piece Canister Set
Options: [Metal', Plastic', Glass', 'Square', 'Wood', ...]

GT Answer: Metal
GPT-4V: Metal
LLaVA-1.5: Metal
Qwen-VL: Metal
DEFLATE: Plastic
LAVIN: Metal

Figure 9: Representative error cases - home domain. (Domain-level analysis: Section 3.2.1; Attribute-level analysis: Section 3.2.2; Comprehensive error analysis and remaining challenges: Appendix C)



(a) Training Set



(b) Evaluation Set

Figure 10: Data distribution of domains, attributes, and attribute values for training and evaluation sets. The full-size version of Figure 3.