# Fashion-related Attribute Value Extraction with Visual Prompting

Anonymous ACL submission

## Abstract

Attribute Value Extraction (AVE) is a crucial technology in e-commerce that enables the identification and extraction of specific product attributes and their corresponding values. While most prior research has focused on directly extracting explicit values from text, this paper introduces a multimodal implicit AVE dataset in the fashion domain, which can generate standardized attribute-value pairs for more effective downstream analysis. Additionally, we propose a step-by-step pipeline that separates the generation of attributes and values, alleviating the model's complexity in understanding the task. In the second step, our visual prompting method directs the model's attention to key regions in the images, thereby improving the accuracy of value extraction. Experimental results demonstrate that our approach outperforms several recent strong baselines, and ablation studies further highlight the effectiveness of each component of our method.

## 1 Introduction

The Attribute Value Extraction (AVE) task aims to automatically identify and extract attributes and their corresponding values from content related to specific items, such as product descriptions, reviews, and social media posts (Li et al., 2021; Khandelwal et al., 2023; Fang et al., 2024). As shown in Figure 1, the input consists of multimodal product information, combining textual descriptions and images. The system processes this data to extract and present attribute-value pairs in the output, such as identifying *"Sleeve Style"* as *"Long Sleeve"* and *"Neckline"* as *"Henley"*.

In this paper, we mainly focus AVE for fashion-related products, and aims to extract the attributes such as size, material, and color, along with their corresponding values (Ahuja et al., 2020). This not only enhances product search efficiency (Nguyen et al., 2020) but also facilitates personalized recommendations (Dong et al., 2020), delivering a better shopping experience for users while aiding platforms in optimizing operations.

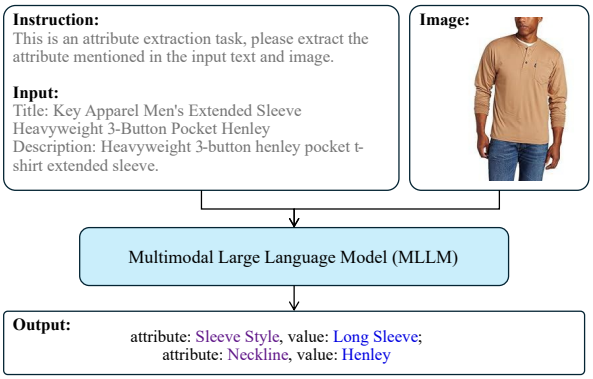


Figure 1: Example of Attribute Value Extraction (AVE).

However, Attribute Value Extraction (AVE) tasks still face several limitations and significant challenges. One of the limitations in recent Attribute Value Extraction (AVE) tasks is the high variability in how values are expressed. The same value can be represented in multiple ways. For instance, *"v-neck"* may appear as *"V-neck"*, *"plunging neckline"* or *"V-shaped neck"*. This complicates subsequent analysis and aggregation. Works such as Zhang et al. (2021); Li et al. (2021); Wang et al. (2020) extract the corresponding values directly from the text without a classification process, which contributes to the diverse expressions of values, reducing the method's overall effectiveness and utility. Another challenge is that images are playing an increasingly important role in conveying product information, making text alone insufficient for accurate attribute extraction. Integrating images is essential for better information extraction, but it presents a challenge due to the complexity and richness of visual content, which makes it harder for models to extract key information.

For addressing challenge of the high variability of values, we construct a new dataset with implicit expression and standardized attribute values.

Specifically, we take two approaches: First, we apply synonym substitution to replace explicitly mentioned values in the input text, enriching the dataset and allowing for more flexible reasoning. Second, we manually standardize the expression of values within the same category, consolidating variations to ensure consistency. This implicit multimodal dataset serves as a valuable benchmark for advancing attribute-value extraction tasks, particularly in the clothing domain.

In addition, we propose a two-step framework to replace the simultaneous generation of attributes and values to address the challenge from multimodal input. In the first step, we provide the input text and image to the model with an instruction to extract attributes, such as *"Sleeve Style"* and *"Neckline"*. In the second step, We ask our model to extract the relevant values according to the attributes extracted in the first step. Beside telling model the known attributes in the form of text, we apply visual prompting (Wu et al., 2024; Yu et al., 2025) and highlight the given attributes in images to reduce the difficulty of the model in visual information extraction.

Our experimental results demonstrate that our proposed method achieves outstanding performance, effectively extracting high-quality attribute-value pairs from multimodal inputs. Additionally, we compare our method with a text-based baseline on both implicit and explicit datasets, revealing that our method yields more significant improvements on implicit data. Furthermore, we evaluate different visual prompting techniques and confirm that our method delivers the best results.

## 2 Related Works

Traditional AVE datasets typically focus on extracting attributes and values from text (Yan et al., 2021; Yang et al., 2022). As images play an increasingly significant role in product information, more and more AVE datasets are emerging in the multimodal domain (Zhu et al., 2020; Zou et al., 2024; Zhang et al., 2023). According to whether the attribute-value pairs are explicitly stated in the text, AVE datasets can be categorized into explicit and implicit datasets. Explicit AVE (Zheng et al., 2018; Xu et al., 2019) involves extracting directly stated attribute-value pairs, while implicit AVE (Zou et al., 2024; Zhang et al., 2023) focuses on inferring attributes and values from the context when they are not explicitly mentioned.

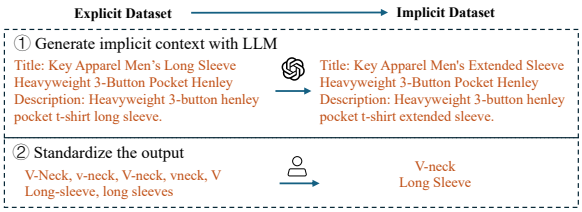


Figure 2: Example of data construction.

AVE tasks can be divided into two types: one uses encoded inputs with the BIO tagging scheme for extraction, typically in NER tasks, while the other uses generative models to produce attribute-value pairs. Kozareva et al. (2016) proposed BiLSTM-CRF, a baseline NER model for attribute-value prediction. Zhu et al. (2020) introduced a multi-modal NER framework, integrating multi-modal features with cross-modality attention for joint attribute prediction. JG-AVE (Roy et al., 2022) tackles the AVE task in a generative framework by formulating it with a word sequence-based paradigm and a positional sequence-based paradigm. These works are done under explicit dataset. With the development of Large Language Model (LLM), generative methods has been more widely used, especially on the implicit datasets. Zou et al. (2024) proposed a multimodal implicit dataset and evaluated it on six recent multimodal large language models (MLLMs) with eleven variants. Fang et al. (2024) proposed a novel algorithm called LLM-ensemble to ensemble different LLMs' outputs for attribute value extraction.

Different from previous studies, We introduce a multimodal implicit dataset that is more representative of real-world scenarios. Moreover, we propose a Step-by-Step pipeline with visual promoting to reduce the difficulty for the model in simultaneously extracting attributes and values, improving extraction accuracy.

## 3 Implicit Dataset Construction

### 3.1 Dataset Construction

In order to get an implicit AVE dataset that has multi-attributes and standardized values, we construct a dataset referred to as *FashionMAVE*, which is based on a subset in fashion domain of the MAVE dataset (Yang et al., 2022), the largest product attribute-value extraction dataset based on the number of attribute-value pairs.

Specifically, we generate the implicit context using GPT4o-mini (OpenAI, 2025), where we pro-

| Dataset | IM | MM | MA | SV |
|---|---|---|---|---|
| OpenTag(Zheng et al., 2018) | ✗ | ✗ | ✓ | ✗ |
| AE-110(Xu et al., 2019) | ✗ | ✗ | ✓ | ✗ |
| MEPAVE(Zhu et al., 2020) | ✗ | ✗ | ✓ | ✗ |
| AdaTag(Yan et al., 2021) | ✗ | ✗ | ✓ | ✗ |
| MAVE(Yang et al., 2022) | ✗ | ✗ | ✓ | ✗ |
| ImplicitAVE(Zou et al., 2024) | ✓ | ✓ | ✗ | ✗ |
| DESIRE(Zhang et al., 2023) | ✓ | ✓ | ✓ | ✗ |
| Ours | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparsion between datasets. In this table, *IM* means implicit dataset, *MM* means: multimodal, *MA* means: multi-attribute and *SV* means standardized value.

| Category | Attribute | Train | Val | Test |
|---|---|---|---|---|
| Shirt | Neckline | 1365 | 151 | 725 |
| | Sleeve Style | 1148 | 125 | 632 |
| | Fastening Style | 152 | 17 | 125 |
| | Pattern | 112 | 12 | 73 |
| | Shoulder Style | 58 | 6 | 58 |
| | Active Style | 33 | 1 | 5 |
| Dress | Sleeve Style | 568 | 87 | 419 |
| | Neckline | 897 | 96 | 501 |
| | Pattern | 175 | 15 | 115 |
| | Length | 806 | 134 | 595 |
| | Shoulder Style | 394 | 35 | 118 |

Table 2: Attribute distribution in our dataset.

| Dataset | Attribute | Explicit | Implicit |
|---|---|---|---|
| Shirt | Neckline | 40 | 21 |
| | Sleeve Style | 40 | 11 |
| | Fastening Style | 14 | 4 |
| | Pattern | 15 | 7 |
| | Shoulder Style | 3 | 3 |
| | Active Style | 4 | 2 |
| Dress | Sleeve Style | 29 | 11 |
| | Neckline | 36 | 18 |
| | Pattern | 24 | 8 |
| | Length | 45 | 13 |
| | Shoulder Style | 8 | 3 |

Table 3: Number of values for each attribute in explicit and implicit dataset.

vide explicit values and ask the model to try to replace these values in the input context with alternative expressions, while preserving the original meaning. As shown in Figure 2, the explicit expression *"long sleeve"* is replaced with *"extended sleeve"* (implicit), while *"Henley"* remains unchanged due to the lack of alternative expressions. This process enables our model to generate results by understanding the input context rather than merely extracting named entities, thus improving its generalization performance in real-world applications.

To further standardize the output, we manually reviewed all values and created a mapping between explicit values and their standardized counterparts. As shown in Figure 2, expressions like *"V-Neck"*, *"v-neck"*, *"V-neck"*, *"vneck"* and *"V"* are standardized to *"V-neck"* while *"Long-sleeve"* and *"long sleeves"* are standardized to *"Long Sleeve"*. This ensures that outputs with the same meaning but different expressions are standardized, making the results more consistent and applicable in real-world applications.

### 3.2 Comparison with Other Datasets

Table 1 provides a comparison between our dataset and others. Traditional AVE datasets are primarily explicit and unimodal (Zheng et al., 2018; Xu et al., 2019; Zhu et al., 2020; Yan et al., 2021; Yang et al., 2022). While they include multiple attributes, they do not capture implicit expressions of values or ensure value standardization. (Zou et al., 2024) introduces a multimodal implicit dataset based on MAVE, but each data point is only paired with a single attribute-value pair. DESIRE (Zhang et al., 2023) annotates values for candidate attributes of each product but does not standardize its annotations. This comparison highlights the advantages of our proposed dataset in effectively handling implicit expressions, supporting multimodal input, incorporating multi-attribute labels, and ensuring value standardization.

### 3.3 Dataset Statistics

In MAVE, all data is explicit, meaning that the value of each data point can be directly found in the text. After implicit processing, the majority of the values in our dataset has become implicit, with explicit data now accounting for only about 2.5%. We ensure that all attributes appearing in our dataset are represented at least ten times in the training set to maintain data balance. Moreover, all attributes present in the test and validation sets have also been observed in the training set, ensuring consistency across all splits.

In our experiment, we split the dataset into three subsets at a ratio of 1800:200:1000, where 1800 samples were allocated to the training set, 200 samples to the validation set, and 1000 samples to the test set. The distribution of different attributes across these sets is shown in Table 2.

We also count the number of values for different

attributes in our dataset (Implicit). Additionally, we performed the same analysis on the original MAVE dataset, which corresponds to our dataset in its explicit form. The results, shown in Table 3, indicate that our dataset successfully reduces the variety of value types by standardizing expressions for values with the same meaning.

# 4 Methods

In this study, our goal is to detect the attribute-value pairs of fashion-related products according to their descriptive text and images. In our task, each attribute has a unique corresponding value while each product can have multiple attributes.

In our task, the process of extracting attribute-value pairs is illustrated in Figure 3. This is a step-by-step procedure: in the first step, a multimodal large language model (MLLM) extracts attributes (e.g., *"sleeve style", "neckline"*) from both text and images. In the second step, the extracted attributes are used to reformulate the input instructions, and visual prompting is applied, using Grounding DINO (Liu et al., 2023b) to detect attribute-related image regions and grayscaling non-target areas to minimize distractions. This process enables the model to focus on key information, ultimately extracting accurate attribute-value pairs (e.g., *"Sleeve Style: Long Sleeve"*). This approach ensures better reasoning and enhances implicit attribute value extraction.

## 4.1 Attribution Extraction

In this step, we utilize both the product's text description and image as input, leveraging a multimodal large language model (MLLM) to extract relevant attributes. This approach enables the model to integrate textual and visual information, improving the accuracy and completeness of attribute extraction. The process is defined as follows:

$$\mathrm{A}[a_1, a_2, ..., a_k] = \mathrm{MLLM}(T_1, I_1) \quad (1)$$

where $\mathrm{A}[a_1, a_2, ..., a_k]$ denotes the extracted attributes, and $T_1$ and $I_1$ represent the input text and image, respectively. For the MLLM architecture, we use LLAVA1.5 (Liu et al., 2023a), which combines CLIP (Radford et al., 2021) for image processing and Vicuna 1.5 (Zheng et al., 2023) for text understanding. CLIP extracts visual features, aligning them with textual representations, while Vicuna 1.5 processes the text descriptions to capture key semantic details. These features are then fused using a Transformer-based module, allowing the model to reason jointly across both modalities and extract the most relevant attributes effectively.

## 4.2 Value Extraction

### 4.2.1 Instruction Design

In the Attribution Extraction step, the multimodal large language model (MLLM) is instructed to extract attributes from both textual descriptions and images. Once the relevant attributes have been identified, we reformulate the instruction in input text to provide clearer guidance for the model in the Value Extraction step. This ensures that the model can focus on generating precise attribute values based on the detected attributes. The new input text is denoted as $T_2$:

$$\mathrm{T}_2 = \mathrm{InstructionDesign}(T_1, A) \quad (2)$$

where $T_1$ and $A$ represent the original input text and the generated attributes, respectively. The design of the instruction can be seen in the input of the second step in Figure 3.

### 4.2.2 Visual Prompting

In this part, we use visual prompting to emphasize the regions in images corresponding to the extracted attributes and make it easier for the model to accurately generate corresponding values according to the attributes.

**Detection of Attributes Related Regions**

We firstly use Grounding DINO (Liu et al., 2023b), an open-set object detection model, to detect these specific areas that correspond to the extracted attributes from raw images. Given the set of attributes $\mathrm{A}[a_1, a_2, \ldots, a_k]$ extracted by the Multimodal Large Language Model (MLLM), Grounding DINO is applied to perform object detection and localize the regions of the attributes in the image. This process is defined as follows:

$$\mathrm{R}[a_1, a_2, \ldots, a_k] = \mathrm{GroundingDINO}(\mathrm{I}_1, \mathrm{A}) \quad (3)$$

Where $\mathrm{R}[a_1, a_2, \ldots, a_k]$ represents the bounding box coordinates of the regions corresponding to each extracted attribute in the image, $\mathrm{I}_1$ denotes the input image, and A is the set of attributes extracted in the previous step. Grounding DINO is responsible for detecting these attributes in the image and outputting the bounding box coordinates for their respective locations. For instance, if the attribute is *"sleeve"*, Grounding DINO outputs the

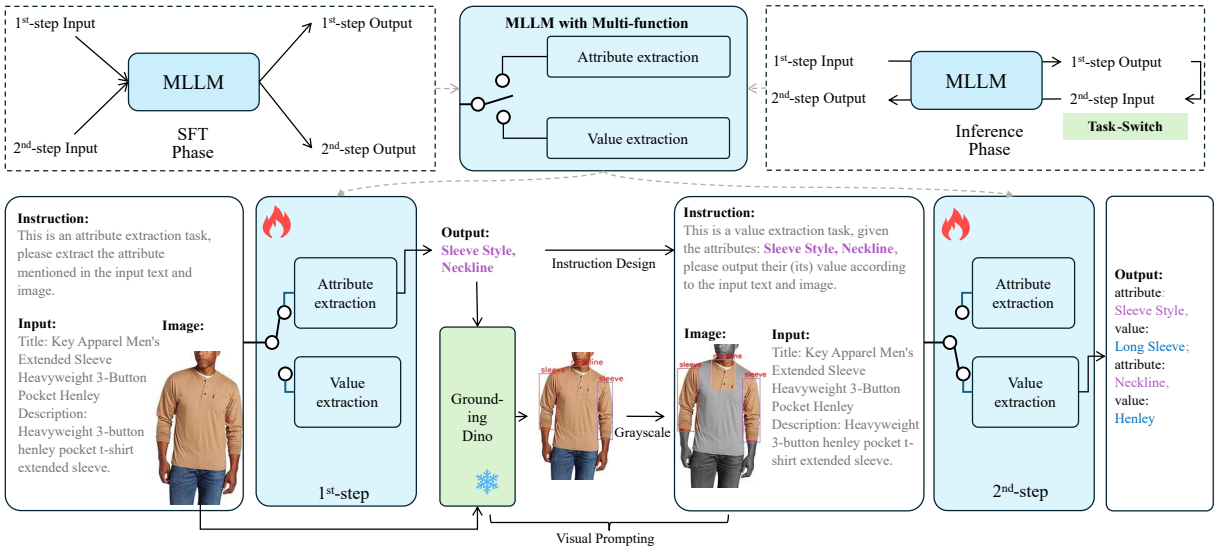


Figure 3: The overview of proposed model.

regions in the image corresponding to the sleeves, enabling precise localization of the attributes.

### Grayscale Attributes Related Regions

To further enhance attribute-value extraction accuracy by reducing the influence of unimportant information and ensuring that the model complies with the most informative visual cues, we converted the image outside the detected attribute areas into grayscale. Specifically, we retain the color information within these regions while converting the rest of the image to grayscale.

$$\mathrm{I}_2 = \mathrm{Gray}(I_1, R) \tag{4}$$

The grayscale procedure is achieved by replacing the RGB values of the non-relevant areas with their corresponding grayscale value. This approach ensures that the model can focus more effectively on the key attributes by suppressing the less important parts of the image, thus improving the accuracy of attribute value generation.

After obtaining the instruction with attributes and images with visual prompting, we generate the final attribute-value pairs through the MLLM, which follows the same process as the Attribution Extraction step:

$$[(a_1, v_1), \ldots, (a_k, v_k)] = \mathrm{MLLM}(T_2, I_2) \tag{5}$$

### 4.3 SFT & Inference Phase

Our step-by-step multimodal framework requires the MLLM to effectively detect both attributes and attribute pairs from datasets of varying formats. To achieve this, the model is trained with a multi-task strategy, enabling it to dynamically switch functions based on the task type. As illustrated in the SFT (supervised fine-tuning) phase of Figure 3, data from both Step-1 and Step-2 is merged and shuffled during training to enhance the model's adaptability.

In the inference phase, the input data is first processed in the format required for Step-1. Recognizing that it is an attribute extraction task, the model identifies and extracts attributes based on the given input. These extracted attributes are then used to generate new instructions and images through visual prompting, transforming the data into the format required for Step-2. This facilitates seamless task-switching, allowing the model to recognize the transition to a value extraction task. Finally, the model generates the corresponding attribute-value pairs.

## 5 Experiments

### 5.1 Experimental Setting

We evaluate our experimental results from two perspectives: *attribute* and *attribute-value pair*. The evaluation of attributes focuses on assessing the model's ability to extract attributes from the dataset, while the evaluation of attribute-value pairs measures the model's performance in correctly identifying values and pairing them with their corresponding attributes. For both perspectives, we use precision, recall, and F1-score as evaluation metrics, with F1-score being the primary metric.

| Method | | Shirt | | | Dress | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| *Performance on Attribute Extraction* | | | | | | | |
| Language Models | T5 | 89.14 | 87.08 | 88.10 | 88.37 | 80.34 | 83.99 |
| | GPT4o-mini | 55.78 | 57.29 | 56.52 | 50.38 | 59.49 | 54.56 |
| | LLaMA3 | 94.43 | 91.22 | 92.80 | 90.60 | 87.64 | 89.09 |
| | InternLM2.5 | 92.53 | 88.65 | 90.55 | 92.34 | 88.76 | 90.51 |
| Multimodal Models | Qwen2-VL | 94.65 | 87.73 | 91.06 | 93.12 | 87.48 | 90.21 |
| | BLIP2 | 90.25 | 87.67 | 88.94 | 91.96 | 88.37 | 90.14 |
| | InternXComposer2 | 92.65 | 88.01 | 90.27 | 88.67 | 87.63 | 87.98 |
| | LLAVA1.5 | 91.76 | 91.64 | 91.70 | 89.31 | 89.47 | 89.39 |
| | DEFLATE | 91.25 | 86.90 | 88.04 | 89.21 | 80.10 | 84.63 |
| | Ours | 93.82 | 94.75 | 94.28 | 91.23 | 90.89 | 91.05 |
| *Performance on Value Extraction* | | | | | | | |
| Language Models | T5 | 66.63 | 73.79 | 69.57 | 65.96 | 72.90 | 69.32 |
| | GPT4o-mini | 25.78 | 28.04 | 26.86 | 19.81 | 22.13 | 20.09 |
| | LLaMA3 | 83.43 | 86.37 | 84.88 | 78.71 | 81.32 | 80.00 |
| | InternLM2.5 | 83.29 | 86.48 | 84.86 | 78.67 | 82.98 | 80.81 |
| Multimodal Models | Qwen2-VL | 82.09 | 88.56 | 85.21 | 78.61 | 83.68 | 81.07 |
| | BLIP2 | 80.54 | 85.15 | 82.80 | 86.58 | 77.35 | 81.64 |
| | InternXComposer2 | 83.72 | 82.30 | 83.02 | 76.44 | 78.69 | 77.60 |
| | LLAVA1.5 | 85.18 | 85.24 | 85.21 | 80.23 | 81.51 | 80.86 |
| | DEFLATE | 69.00 | 73.24 | 71.05 | 67.68 | 73.23 | 70.35 |
| | Ours | 89.04 | 85.99 | 87.48 | 82.20 | 82.92 | 82.55 |

Table 4: Comparison with baselines.

For each experiment, we conduct a hyperparameter sweep across learning rate, batch size, and training epochs, selecting the parameters that achieved the highest validation performance. All experiments were performed on a single 4090 GPU. To fine-tune the large language models (LLMs) within the limited memory space of our GPU, we adopted the Low-Rank Adaptation (LoRA) fine-tuning approach (Hu et al., 2021).

### 5.2 Main Results

To thoroughly evaluate the performance of our method, we select several strong baselines and a recent SOTA method for comparison. Specifically, we assess a range of language models, including T5 (Raffel et al., 2020) (base), LLaMA3 (Dubey et al., 2024) (7B), InternLM2.5 (Cai et al., 2024) (7B), and GPT-4o-mini (OpenAI, 2025). Among these, T5 is fully fine-tuned, LLaMA3 and InternLM2.5 are fine-tuned using LoRA, and GPT-4o-mini is evaluated in a zero-shot setting. Additionally, we conduct experiments with multimodal large language models known for their strong performance, including Qwen2-VL (Wang et al., 2024), InternXComposer2 (Dong et al., 2024), BLIP2 (Li et al., 2023) and LLAVA1.5 (Liu et al., 2023a). All of these multimodal models are 7B variants and fine-tuned using LoRA. We also conduct experiments on DEFLATE (Zhang et al., 2023), which is a SOTA method for implicit AVE task.

Based on the results shown in Table 4, it is evident that fine-tuned models consistently outperform the zero-shot model (GPT4o-mini) in both attribute and value extraction tasks. Additionally, Large Language Models (LLMs) demonstrate superior performance compared to T5, which has relatively fewer parameters. Interestingly, despite their multimodal capabilities, multimodal Large Language Models do not consistently surpass their unimodal counterparts, particularly when compared to LLaMA3 and InternLM2.5. This observation highlights that simply incorporating multimodal inputs may not fully capture or exploit the rich information present in images. Our proposed method, which integrates step-by-step reasoning with visual prompting, effectively reduces reasoning complexity at each step. As a result, it achieves state-of-the-art performance in both attribute and value extraction, demonstrating the effectiveness of our method in addressing the challenges of this task.

### 5.3 Ablation Studies

To validate the effectiveness of our proposed method, we conduct a series of ablation studies. These experiments include: using text only, using both text and images for direct output, and employing a basic Step by Step approach without visual prompting. The results (F1-Score) are shown in

| | Shirt | | Dress | |
|---|---|---|---|---|
| | Attr. | Val. | Attr. | Val. |
| Ours | 94.28 | 87.48 | 91.05 | 82.55 |
| w/o V-Prompting | 93.35 | 85.31 | 90.63 | 81.11 |
| w/o Step-by-Step | 91.70 | 85.21 | 89.39 | 80.86 |
| w/o Image | 92.33 | 84.93 | 88.70 | 80.08 |

Table 5: Ablation studies. *V-Prompting* means visual prompting

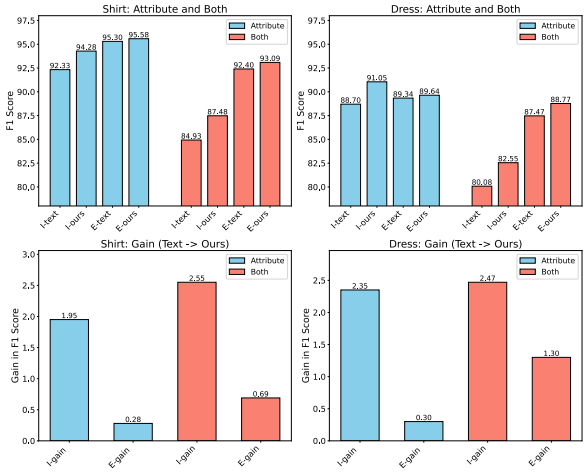


Figure 4: Comparison between explicit and implicit dataset. *I* represents experiments on implicit dataset and *E* represents experiments on explicit dataset, *text* means using text as only input and *Ours* means using our proposed method.

Table 5.

The results show that using text and images for direct output with LLAVA1.5 does not yield significant improvement over using text alone. This suggests that the simple inclusion of images in this experiment provides limited contributions. However, when we incorporate step-by-step inference (without visual prompting), we observe an improvement in the extraction of attributes and attribute-value pairs. This indicates that step-by-step inference effectively reduces the complexity of model inference and enhances extraction performance. Finally, our proposed method further improves upon the basic step-by-step approach by adding visual prompting, both in terms of attributes and attribute-value pair extraction. This demonstrates that the visual prompting in our method helps the model more easily extract values based on the given attributes. Moreover, while the first step of our method is similar to basic step-by-step method, the more focused inference in the second step enables the model to better understand the attributes, leading to improved performance in attribute extraction.

### 5.4 Comparison between Explicit and Implicit Dataset

To further demonstrate the efficiency of our method, we conducted experiments on the explicit dataset corresponding to our implicit dataset. These experiments were carried out using our proposed method and compared against a baseline that utilizes text input only.

The experimental results (F1-Score) are provided in the above two sub-figure of Figure 4, which highlight the effectiveness of our proposed method across both implicit and explicit datasets for fashion-related attribute-value extraction tasks.

When comparing the explicit and implicit datasets, the explicit dataset demonstrates better performance. This is because the outputs of the explicit dataset are directly derived from the input text, making it easier for the model to detect values without requiring extensive analysis or inference.

The below two sub-figures in Figure 4 show the performance gain compared with our method and base method. In these figures we observe smaller improvement achieved in the explicit dataset when employing our method. This may be due to the fact that, in explicit datasets, AVE tasks are more akin to Named Entity Recognition (NER), which primarily focuses on extracting named entities from text and places less emphasis on utilizing image information. These experiments confirm that our proposed method is effective on both explicit and implicit datasets, with its potential being more fully realized in implicit datasets, where reasoning and multimodal analysis play a more critical role.

### 5.5 Influence of Visual Prompting Methods

We also conduct experiments on various visual prompting methods. Figure 5 (a) shows the original images without any visual prompting. The first method, shown in Figure 5 (b), highlights the attributes in boxes with labels above. This represents a simplified version of the approach we employed. The result (F1-Score) of this method are presented in Table 6, which shows that it performs less effectively compared to our proposed method. This indicates that the grayscale processing applied to non-key areas (Figure 5(c)) successfully reduces the difficulty for the model during inference, leading to an improvement in its performance.

Another method, applied to the API (Yu et al.,

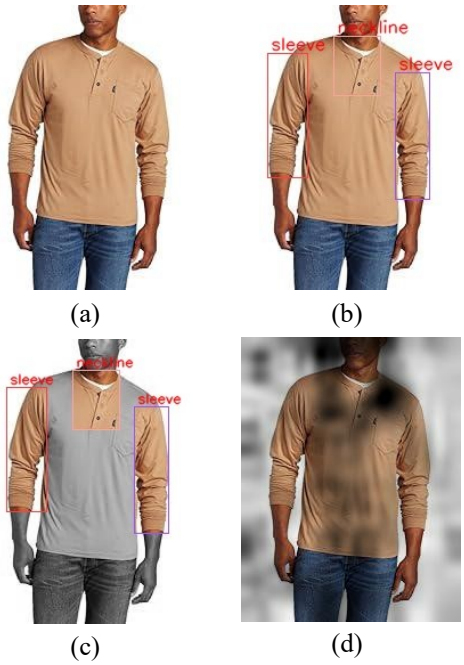


Figure 5: Examples of different visual prompting.

| | Shirt | | Dress | |
|---|---|---|---|---|
| | Attr. | Val. | Attr. | Val. |
| (a)Image only | 91.70 | 85.21 | 89.39 | 80.86 |
| (b)Image+Box | 93.97 | 86.34 | 90.48 | 81.64 |
| (d)Image+API | 92.50 | 86.01 | 88.94 | 80.20 |
| (c)Ours | 94.28 | 87.48 | 91.05 | 82.55 |

Table 6: Results of different visual prompting.

2024), incorporates textual information into images by prompting the model to highlight specific regions of the image based on the provided textual descriptions. The processed image is shown in Figure 5 (d). From the figure, we can observe that the regions around the sleeves and neckline are minimally masked, indicating that this method somewhat achieves its intended effect. However, it is apparent that the mask processing of non-critical areas does not effectively emphasize the crucial parts of the image, leading to a vague and imprecise annotation. The results in Table 6 also suggest that this is not the most effective visual prompting method for our task.

### 5.6 Case Studies

In Figure 6, we present a case study comparing our method to baselines that simply use both text and multimodal input. In the first case, when using text-only input, the model correctly predicts *"long sleeve"* from *"extended-sleeve"* in the text. However, when an image is added, the rolled-up sleeves in the image mislead the model, resulting in the incorrect prediction of *"short sleeve"*. Our method, which employs visual prompting, helps the model focus more on the relevant sleeve details, enabling it to make the correct prediction. In the second case, the term *"Open"* in the input text is ambiguous, as it could refer to either *"off shoulder"* or *"cold shoulder"*. This ambiguity leads the model to make an incorrect prediction regarding the shoulder style. Both our method and the multimodal baseline leverage the additional information from the image to correctly resolve this ambiguity and make accurate prediction.

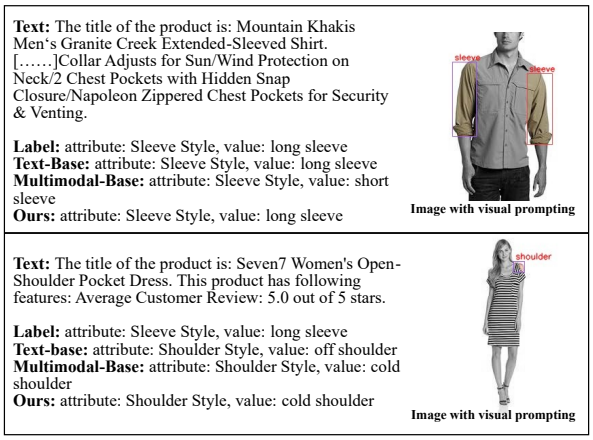


Figure 6: Examples of case study.

Based on the examples above, our method enhances the model's visual reasoning ability, allowing it to focus on key image details while avoiding irrelevant noise. It also resolves text ambiguities by leveraging image information, achieving precise multimodal inference for accurate attribute-value extraction.

## 6 Conclusion

In this paper, we proposed a method that enhances multimodal attribute-value extraction for fashion-related products. By combining step-by-step inference and visual prompting, our approach improves the model's performance in process multimodal input. Our experiments demonstrate that our method outperforms both text-only and multimodal baselines, effectively resolving text ambiguities and focusing on relevant visual details. This work contributes to more accurate and robust attribute-value extraction in the fashion domain, offering potential for improved e-commerce product categorization and recommendation systems.

## 7 Limitations

A limitation of our approach is that it relies on visual cues, making it effective for attributes that can be seen, like sleeve style. However, it is less suitable for extracting intrinsic attributes, such as fabric material, which cannot be easily determined from images. This limits the method's applicability in certain domains.

## References


Aman Ahuja, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan K Reddy. 2020. Language-agnostic representation learning for product search on e-commerce platforms. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 7–15.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Xin Luna Dong, Xiang He, Andrey Kan, Xian Li, Yan Liang, Jun Ma, Yifan Ethan Xu, Chenwei Zhang, Tong Zhao, Gabriel Blanco Saldana, et al. 2020. Autoknow: Self-driving knowledge collection for products of thousands of types. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2724–2734.

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*.

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. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2910–2914.

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

Anant Khandelwal, Happy Mittal, Shreyas Sunil Kulkarni, and Deepak Gupta. 2023. Large scale generative multimodal attribute extraction for e-commerce attributes. *arXiv preprint arXiv:2306.00379*.

Zornitsa Kozareva, Qi Li, Ke Zhai, and Weiwei Guo. 2016. Recognizing salient entities in shopping queries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

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.

Yang Li, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2021. Mave: A product dataset for multi-source attribute value extraction. *Cornell University - arXiv,Cornell University - arXiv*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Thanh Nguyen, Nikhil Rao, and Karthik Subbian. 2020. Learning robust models for e-commerce product search. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6861–6869, Online. Association for Computational Linguistics.

OpenAI. 2025. Chatgpt-4 mini. `https://www.openai.com/`. Version 4.0, AI language model.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Kalyani Roy, Tapas Nayak, and Pawan Goyal. 2022. Exploring generative models for joint attribute value extraction from product titles. *arXiv preprint arXiv:2208.07130*.

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

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*, pages 47–55.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. 2024. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

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, 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.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024. Api: Attention prompting on image for large vision-language models.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2025. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pages 251–268. Springer.

Danqing Zhang, Zheng Li, Tianyu Cao, Chen Luo, Tony Wu, Hanqing Lu, Yiwei Song, Bing Yin, Tuo Zhao, and Qiang Yang. 2021. Queaco: Borrowing treasures from weakly-labeled behavior data for query attribute value extraction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4362–4372.

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.

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*, pages 1049–1058.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

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*.

Henry Peng Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip S Yu, and Cornelia Caragea. 2024. Implicitave: An open-source dataset and multimodal llms benchmark for implicit attribute value extraction. *arXiv preprint arXiv:2404.15592*.

## A Example Appendix

This is an appendix.