









MuJo-SF: Multimodal Joint Slot Filling for Attribute Value Prediction of E-commerce Commodities

Meihuizi Jia , Lei Shen , Luu Anh Tuan , Meng Chen , Jing Xu ,
Lejian Liao , Shaozu Yuan , and Xiaodong He , *Fellow, IEEE*

Abstract—Supplementing product attribute information is a critical step for E-commerce platforms, which further benefits various downstream tasks, including product recommendation, product search, and product knowledge graph construction. Intuitively, the visual information available on e-commerce platforms can effectively function as a primary source for certain product attributes. However, existing works either extract attribute values solely from textual product descriptions or leverage limited visual information (e.g., image features or optical character recognition tokens) to assist extraction, without mining the fine-grained visual cues linked with the products effectively. In this paper, we propose a novel task - Multimodal Joint Slot Filling (MuJo-SF) - that aims to combine multimodal information from both product descriptions and their corresponding product images to jointly fill values into the pre-defined product attribute set. To this end, we develop MAVP, a new dataset with 79k instances of product description-image pairs. Specifically, we present a strategy to fulfill visualized saliency ascription, which aims to distinguish between text-dependent and image-dependent attributes. For those image-dependent attributes, we annotate the corresponding values from images using distant supervision. Then, we design a model for MuJo-SF, which combines multimodal representations and fills image-dependent and text-dependent attributes separately. Finally, we conduct extensive experiments on MAVP and provide rich results for MuJo-SF, which can be used as baselines to facilitate future research.

Index Terms—Attribute value prediction, multimodal joint slot filling, distant supervision, e-commerce commodity

I. INTRODUCTION

AS customers navigate the e-commerce product pages to make product selections, they partake in a multimodal experience, perusing through both product descriptions and images. In this context, customers have the ability to extract attribute values from these multimodal product sources. These attribute values play a pivotal role in assisting customers to identify products that align with their requirements and facilitate their purchasing choices.

This work is supported by the National Key R&D Program of China under Grant No. 2020AAA0106600 and the China Scholarship Council (CSC), Grant No. 202006030076. (Corresponding author: Meng Chen and Lejian Liao.)

Meihuizi Jia is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, 100081. She is also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798. (E-mail: jmhuzi24@bit.edu.cn).

Luu Anh Tuan is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798. (E-mail: anhtuan.luu@ntu.edu.sg).

Lei Shen, Meng Chen, Shaozu Yuan and Xiaodong He are with the JD AI Research, Beijing, China, 100176. (E-mail: shenlei20@jd.com; chenmengdx@gmail.com; yuanshaozu@jd.com; xiaodong.he@jd.com).

Jing Xu and Lejian Liao are with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, 100081. (E-mail: xujing@bit.edu.cn; liaolj@bit.edu.cn).

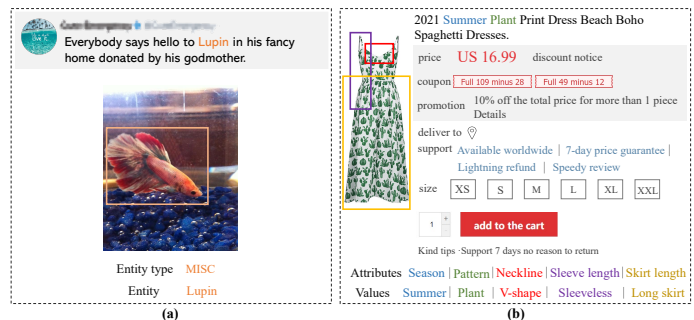


Fig. 1: A comparison of multimodal information extraction (a) in social media scenario (b) in e-commerce scenario.

Most existing methods acquire product attribute values in a text-only landscape, which extract information from the product title and descriptions [1]–[5]. However, product images available on e-commerce platforms can provide substantial contribution in product attribute extraction. For instance, during the process of browsing an e-commerce webpage and seeking a black short-sleeve shirt, presenting the product image can yield instantaneous results as opposed to sifting through numerous pages of product descriptions with the expectation of encountering attributes like color and sleeve length, which the retailer may or may not have included on the product page. Some other methods focus on multimodal attribute value extraction, but they either use image features to supplement textual information [6]–[8] or use optical character recognition (OCR) technology to capture text in the image and correct the words in the product description [9], [10], lacking of fine-grained analysis and mining for product images, and attribute value extraction is inclined to rely on textual information flow.

Figure 1 illustrates the difference in multimodal information extraction between e-commerce and social media platforms. For example, in a social media scenario like Twitter, some named entities of pre-defined types need to be mined (such as persons, locations, organizations and miscellaneous). Images are usually used as auxiliary information of text to alleviate ambiguity in natural languages. As shown in Figure 1(a), the region with the goldfish in the image can be used to assist in recognizing “Lupin” in the Tweet as *miscellaneous* rather than *persons*. But, the goldfish’s name, “Lupin”, cannot be directly obtained from the image. However, in an e-commerce scenario, the product attribute values that need to be mined are typically generic product information. Images should not be merely employed as supplementary information to text; rather, they constitute

the principal source for certain attribute values. As shown in Figure 1(b), the product description includes the value “*Summer*” corresponding to attribute *Season*, as well as the value “*Plant*” corresponding to attribute *Pattern*. However, we can easily access additional product attribute values by paying attention to the product image, such as value “*V-shape*” corresponding to attribute *Neckline*, value “*Sleeveless*” corresponding to attribute *Sleeve length*.

In this paper, we propose a novel task named **Multimodal Joint Slot Filling for Attribute Value Prediction of E-commerce Commodities (MuJo-SF)**, which differs from traditional information extraction systems. In MuJo-SF, the visual information available on e-commerce platforms can serve as a primary source for certain product attributes. Attribute slots for each product category on e-commerce platforms are predefined, and models are required to simultaneously populate all these slots. This necessitates the utilization of information streams from both textual and visual modalities in a collaborative manner. This new paradigm contains three advantages: (1) It simulates human intuition when browsing e-commerce pages and can fully take advantage of different data resources on e-commerce platforms. (2) Models are relieved from predicting the product’s attribute types in advance. (3) Models are expected to accurately and completely fill product slots across all categories, which involves not only predicting attribute values within the same category but also distinguishing attributes across different categories.

To address MuJo-SF, we contribute a new **Multimodal Attribute Value Prediction dataset (MAVP)** consisting of 79k product description-image instances and 25 slots across 5 domains. We first design a visualized saliency ascription strategy to conduct a fine-grained analysis on images, thereby determining which attributes can be directly and easily extracted from the images. Specifically, we mask out the token in the product description that corresponds to the value and then predict this masked token using both textual and multimodal information separately. Based on the ascription results, we categorize all attribute slots into either text-dependent or image-dependent attribute slots. Subsequently, for the image-dependent attributes, we annotate their corresponding values within product images using the multi-dataset associative distant supervision approach.

We develop a new training paradigm for the new task MuJo-SF, which distinguishes between text-dependent and image-dependent modules. After encoding text and images with their respective encoders, we use a cross-modal interaction module to merge the information flows from different modalities. During the training process, we adopt a joint training strategy to simultaneously optimize both text-dependent and image-dependent modules. By fusing image and text information before dividing into sub-modules for joint training, the model benefits from an integrated multimodal context while allowing each sub-module to specialize in its respective domain. This process enhances the model’s ability to extract and utilize the unique features of each modality, leading to a richer and more accurate performance. We conduct extensive experiments on MAVP and achieves promising results on two evaluation metrics.

In summary, MuJo-SF merges multimodal data to produce richer attribute information, thereby presenting a more detailed product view. This richer attribute information can refine search and filtering capabilities, which guides users to their ideal products and support more informed purchase decisions. Simultaneously, automating the process of filling in attribute values from multiple sources can significantly reduce the need for manual intervention and increase efficiency. The dataset and baseline model are publicly available at <https://github.com/jmhz24/MuJo-SF>.

The contribution of this paper is three-fold:

- We introduce MuJo-SF, a **novel task** that is required to utilize information streams from both textual and visual modalities in a collaborative manner to fill all pre-defined product attribute sets. This task simulates human browsing habits on e-commerce platforms.
- We contribute MAVP, a **new dataset** consisting of 79k product description-image pairs and 25 slots across 5 domains. We design a visualized saliency ascription strategy to distinguish between text-dependent and image-dependent attribute slots and annotate image-dependent attributes in images via multi-dataset associative distant supervision.
- We develop a **new training paradigm** for the novel task MuJo-SF, which distinguishes between text-dependent and image-dependent modules. Benchmarking the MAVP dataset with this paradigm, which would serve as a strong baseline, our empirical results show its potential to stimulate future research in the field of multimodal e-commerce product attribute value prediction.

II. RELATED WORK

A. Attribute Value Extraction

Attribute Value Extraction (AVE) is a subtask within the broader domain of information extraction in Natural Language Processing (NLP) [11]–[14] that focuses on identifying and extracting specific attributes and their corresponding values from unstructured text. Existing methods on AVE mainly fall into two categories. One is the sequence labeling based approach, which defines a label sequence for each product description with the BIO tagging schema [15] and predicts the correct label for each token in the input sequence [16]–[19]. OpenTag [1] proposes a BiLSTM-SelfAttention-CRF architecture to fulfill the MAVE task, which does not use any dictionary or hand-crafted features. [2] adopts one set of BIO tags for any attribute to solve the model explosion and explicitly model semantic representations between attributes and titles using the attention mechanism. AdaTag [3] proposes a multi-attribute value extraction model with an adaptive CRF-based decoder to promote knowledge sharing across multiple attributes. TXtract [4] proposes a taxonomy-aware knowledge extraction model and jointly extracts attribute values and predicts the product’s categories using a multi-task learning strategy. The other mainstream method is to formulate many NLP tasks into the question answering framework, such as named entity recognition [20]–[22], entity relation extraction [23], and sentiment analysis [24]. AVEQA [5] extracts attribute

values via question answering, in which a distilled masked language model is proposed to enhance generalization and a classifier with no-answer capability is developed to deal with the no-answer case. MAVEQA [25] also formulates the attribute value extraction task as a question answering problem, but this work emphasizes efficient modeling of products with structured and long profiles.

Compared with the task of text-based attribute value extraction, our MuJo-SF task, which combines data from multiple modalities, leads to a richer set of product attributes and reduces the potential for inaccuracies that can arise from relying on a single data source.

B. Multimodal Attribute Value Extraction

Multimodal approaches, integrating information from diverse modalities, have drawn substantial interest in the research community [26]–[30]. The Internet encompasses a variety of information sources, such as text, images and videos, providing crucial support for multimodal research [31]–[34]. E-commerce platforms boast abundant product images, thus making the integration of visual information into attribute value extraction is logical. M-JAVE [7] jointly models the attribute prediction and value extraction tasks towards the interactions between attributes and values and extracts values from textual product descriptions with the help of auxiliary product images. PAM [9] and SMARTAVE [10] use optical character recognition (OCR) technology to capture text in the image and correct the words in the product description. The former utilizes a transformer-based sequence-to-sequence model to merge product descriptions, OCR tokens, and visual objects in the product image. The latter designs a structured attention mechanism among hyper-tokens and local-tokens to learn valid product representations. However, above methods treat images as supplementary, aiding in the extraction of attribute values that are derived solely from the text information stream. In contrast, our task involves a multimodal attribute value joint filling process that requires information to be extracted separately from both text and image streams before jointly filling the attribute set. MAE [6] and EKE-MMRC [8] employ a generative approach to obtain attribute values. The former jointly embeds the query, text, and images into a common latent space, combines these embedded vectors using a fusion module and produces the final value prediction via a value decoder. The latter extracts e-commerce knowledge via machine reading comprehension, which encodes questions with multi-modal descriptions as a fusion vector and generates an answer from the fusion vector. However, both works encode information flows from different modalities and fuse them into a fusion vector, which is then fed into the decoder. Different from these two papers, our paper highlights the importance of images. We perform a fine-grained analysis of visual information and enhance the model’s capability in directly extracting image-dependent attributes, such as color and sleeve length, from images.

The multimodal attribute values not only summarize product information to facilitate product searches for customers but also provide fine-grained information for vision-language pre-training models designed for e-commerce contexts. FashionSAP

[35] and FashionViL [36] both focus on fashion-focused vision-language pre-training model on e-commerce platforms. The former proposes a fine-grained vision-language pre-training model based on fashion symbols and attributes prompt. The latter proposes a novel vision-language representation learning framework that includes two innovative fashion-specific pre-training tasks: a multi-view contrastive learning task and a pseudo-attributes classification task.

III. MULTIMODAL JOINT SLOT FILLING

A. Task Definition

In this section, we define the new task MuJo-SF. Specifically, we use $P = \{p_1, p_2, \dots, p_N\}$ to represent a product category set containing N product categories, and $X = \{(D_1, I_1), (D_2, I_2), \dots, (D_M, I_M)\}$ to denote a set of multimodal product sources on the e-commerce platform, where D_M and I_M denote M product descriptions and images, respectively. We pre-define a set of attribute slots $S_i = \{s_1, s_2, \dots, s_{f_i}\}$ for each product category in P , where S_i denotes the attribute slot set for product category p_i , f_i denotes the number of slots of p_i . Common attribute slots are identifiable across diverse product categories, alongside distinct attribute slots that are specific to individual categories. For example, product categories “Clothes” and “Pants” have common attribute slots *Color*, *Material*, *Season* and so on, and they also have unique attribute slots *Sleeve Length* and *Pant Length*, respectively. Thus we merge these common attribute slots and provide a set of attribute slots $S = \{s_1, s_2, \dots, s_F\}$ for all products, where F indicates the number of slots for all product categories. Each product attribute slot corresponds to a set of values, so we define a value set $V = \{V_1, V_2, \dots, V_F\}$ in which each element is a value candidate set corresponding to the slot. For example, $V_i = \{v_1^i, v_2^i, \dots, v_{m_i}^i\}$ denotes the set corresponding to slot s_i , which consists of m_i value candidates. The goal of MuJo-SF is to jointly fill all attribute slots in S via combining multimodal product descriptions and images. Note that if a slot is not mentioned in the multimodal product source, its value will be annotated as “none”.

B. Data Collection

Innovative research in the e-commerce domain benefits from the construction of diverse data resources within these platforms, among which are datasets specifically designed to facilitate the improvement of attribute value extraction [2], [3], [6], [7], [25], [37], [38]. However, to facilitate the research of multimodal joint slot filling, we develop a new multimodal attribute value prediction dataset (named MAVP) which is annotated and re-constructed based on the dataset MEPAVE [7]. MEPAVE is built for multimodal attribute value extraction (MAVE), and the instances in it are collected from a mainstream Chinese e-commerce platform¹, which consists of textual product descriptions and product images. We first re-construct MEPAVE to initialize our new dataset MAVP. We select 5 product categories from the fashion domain and 25 product attributes from MEPAVE, which are *Clothes*, *Pants*,

¹<https://www.jd.com/>

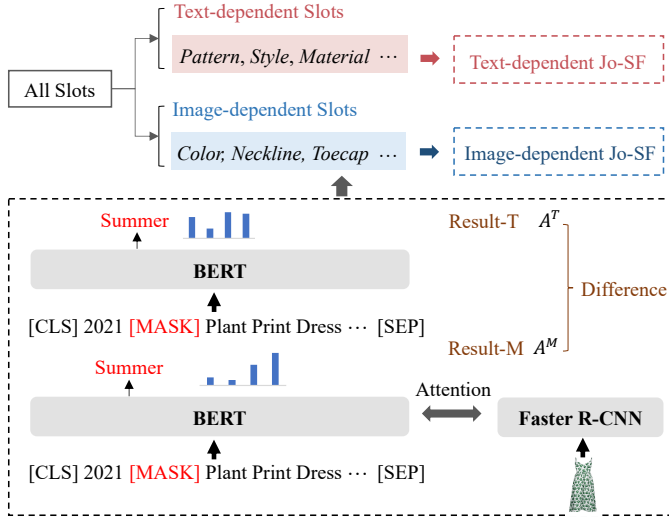


Fig. 2: Process of visualized saliency ascription (Result-T and Result-M denote the result of masked value prediction on textual and multimodal information, respectively).

Dresses, Shoes, and Boots. We design 25 slots corresponding to product attributes. In MEPAVE, all attribute annotations are covered in the product descriptions, product images are usually used as auxiliary information, and there is a lack of annotations involved in them. Next, we devise a strategy for visualized saliency ascription to analyze product images in MEPAVE, subsequently identifying which attributes can be seamlessly and directly extracted from the images. Then we annotate these attributes by using a multi-dataset associative distant supervision approach. Finally, our dataset MAVP can be updated by combining annotations from product descriptions in MEPAVE with the annotations obtained from product images based on ascription strategy. Additionally, we further analyze the quality of annotations with the distant supervision method.

1) *Visualized Saliency Ascription:* As mentioned in the introduction, it is effortless to acquire visualized information like sleeve length and color from product images, while it is challenging to obtain details such as size and elasticity from the images. With the above intuitions, one question arises: Can we estimate visualized information quantitatively? For the question, inspired by [39], we introduce a new concept *visualized saliency ascription*, which can explicitly quantify the contribution of the product image to MuJo-SF. Specifically, there is a set of labeled product descriptions and unlabeled images. We intend to explore the visualized saliency by predicting the values that are annotated in product descriptions with textual and multimodal information, respectively. Next, we depict the visualized saliency ascription strategy, which consists of three steps: (1) Masked value prediction on textual information, (2) Masked value prediction on multimodal information, and (3) Visualized saliency estimation.

Masked Value Prediction on Textual Information. We first denote a product description with k words $D = \{w_1, w_2, \dots, w_k\}$ which contains the value, $v_i \in D$ (for clarity, we remove the subscript of D). We assume the value’s token is w_i . Inspired by the method of “masked language

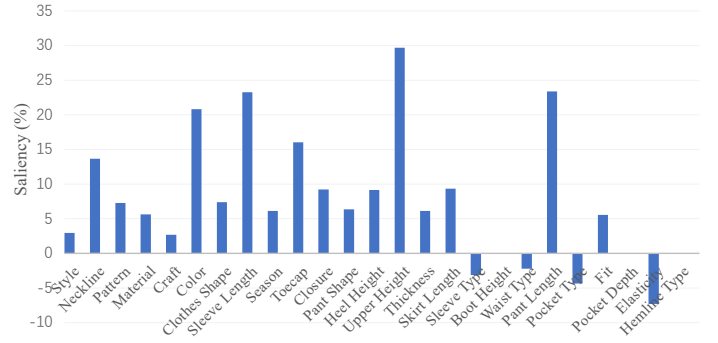


Fig. 3: Illustration of visualized saliency ascription on all slots where the visualized saliency score of one attribute slot can quantify its contribution to our MuJo-SF (Slots without the blue column indicate that the difference is 0).

modeling” (MLM) [40], we replace the value’s token w_i in the product description D with a placeholder [MASK] and adopt the self-supervision method to predict the [MASK] value. Specifically, we employ the pre-trained BERT model [40] as encoder, and feed an input through BERT by concatenating $\{[CLS], w_1, w_2, \dots, [MASK], \dots, w_k, [SEP]\}$, where [CLS] and [SEP] are special tokens. The output of BERT is a contextual representation matrix $\mathbf{H} \in \mathbb{R}^{(k+2) \times d_t}$. We use $\mathbf{h}_w^T \in \mathbb{R}^{d_t}$ to denote the representation of the specific [MASK] token. The masked token is fed into an output softmax over all values and the probability is $P_w^T \in \mathbb{R}^{1 \times Z}$, where Z represents the number of all values. At last, the value’s token w_i is predicted via minimizing the cross-entropy loss function.

Masked Value Prediction on Multimodal Information. In this section, we will combine the product description and the product image to predict the masked value’s token w_i in the product description D . Specifically, we use Faster R-CNN in conjunction with the ResNet-101 [41] to detect objects of the product image and get a set of image features $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_l\}$ [42], $\mathbf{U} \in \mathbb{R}^{l \times d_U}$ where l denotes the number of objects. Next, after receiving the masked token’s representation \mathbf{h}_w^T and image features \mathbf{U} , we use a linear projection to map them to the same dimension d and fuse the textual and visual representations via calculating the attention scores between \mathbf{h}_w^T and \mathbf{U} :

$$\alpha = \text{softmax} \left(\frac{(\mathbf{U}\mathbf{W}_U)(\mathbf{h}_w^T\mathbf{W}_w^T)}{\sqrt{d_h}} \right) \quad (1)$$

$$\mathbf{U}' = \sum_{q=1}^l \alpha_q * \mathbf{u}_q \quad (2)$$

$$\mathbf{h}_w^M = (\mathbf{h}_w^T\mathbf{W}_w^T) * (\mathbf{U}'\mathbf{W}_U) \quad (3)$$

where $\mathbf{W}_U \in \mathbb{R}^{d_U \times d}$, $\mathbf{W}_w^T \in \mathbb{R}^{d_t \times d}$, $\mathbf{U}' \in \mathbb{R}^{d_U}$, $\mathbf{h}_w^M \in \mathbb{R}^d$ represents the fusion representation of the masked value’s token and image, $*$ represents element-wise multiplication. The multimodal representation \mathbf{h}_w^M is fed into an output softmax over all values and the probability is $P_w^M \in \mathbb{R}^{1 \times Z}$. At last, the value’s token w_i is predicted via minimizing the cross-entropy loss function.

Visualized Saliency Estimation. In this part, we quantitatively estimate the visualized saliency score based on the above prediction results P_w^T and P_w^M . Specifically, in Section III-A, we define the attribute slots set $S = \{s_1, s_2, \dots, s_F\}$, and each attribute slot s_i corresponds to a value candidate set V_i . We first count respectively the accuracy of each slot using different information sources (textual or multimodal information) based on value prediction results P_w^T and P_w^M . We define $A^T = \{a_1^T, a_2^T, \dots, a_F^T\}$ and $A^M = \{a_1^M, a_2^M, \dots, a_F^M\}$, where A^T and A^M represent accuracy sets of textual information and multimodal information slots, respectively. For $\forall a_i^T \in A^T$, if $w_i \in V_i \wedge \text{argmax}(P_{w_i}^T) = Y_{w_i}$, then $a_i^T := a_i^T + 1$, where Y_{w_i} denotes the ground truth and w_i involves the value’s token corresponding to the i -th slot. The same updated approach applies to A^M . The visualized saliency scores $Saliency = \{dif(s_1), \dots, dif(s_F)\}$ can be explicitly quantified via calculating the difference between A^T and A^M for attribute slots set S , where $dif(s_i) = a_i^M - a_i^T$.

The two steps of Masked Value Prediction on Textual Information and Masked Value Prediction on Multimodal Information are trained separately. We use Accuracy as evaluation metrics. The *Saliency* is calculated on the validation set of the MEPAVE dataset [7]. For more consistent outcomes, we carry out three experiments with varying seeds and compute the mean results. The results are presented in Figure 3. Based on the results, we divide all attribute slots into an image-dependent set $S_{img} = \{s_i | dif(s_i) \geq \lambda\}$ and a text-dependent set $S_{text} = \{s_i | dif(s_i) < \lambda\}$. The threshold λ is empirically determined. We assume that there are $|S_{img}|$ image-dependent slots and $|S_{text}|$ text-dependent slots in all F slots. The process of Section III-B1 is shown in the Figure 2.

2) *Multi-dataset Associative Distant Supervisor*: In this section, we annotate images in MEPAVE. Benefiting from the visualized saliency ascription for image information in Section III-B1, we annotate the attributes in the image-dependent set S_{img} . In this paper, we empirically determine $\lambda = 9$ and divide the 25 attribute slots into 9 image-dependent attribute slots and 16 text-dependent attribute slots. These 9 image-dependent attribute slots are: *Neckline*, *Color*, *Sleeve Length*, *Toecap*, *Closure*, *Heel Height*, *Upper Height*, *Skirt Length*, *Pant Length*. We propose a multi-dataset associative distant supervision approach to annotate images in MEPAVE. Specifically, we first select three fashion datasets that share the same domain as MEPAVE, and then, we crawl some images from the web to facilitate distant annotation. Attributes possessing comparable meanings yet divergent terminologies or linguistics are systematically standardized into distinct expressions (e.g., mapping “*Sleeveless*” into “*无袖*”). We will introduce these four datasets and the experiments separately. The pre-trained image classification model, Swin-Transformer [43], is utilized to conduct all experiments. The evaluation metric of all experiments in this section is Accuracy. We randomly split the data into training, validation, and test sets in the ratio of 8:1:1 for FashionAI Attribute, UT-ZAP50K, and Crawl Images datasets.

FashionAI Attribute [44]. This is a large-scale attribute dataset with high-quality manual annotation. The dataset is used in a global fashion challenge, in which only single-subject (single-

TABLE I: Statistics and results on FashionAI Attribute Dataset (F-I denotes FashionAI Attribute Dataset, #imgs denotes the number of images, Neck-I denotes Neckline, and R1 and R2 denote Round-1 and Round-2, respectively.).

F-I	Tasks	Length				Design			
		Sleeve	Pant	Skirt	Coat	Collar	Lapel	Neck	Neck-I
R1	#imgs	13,299	7,460	9,223	11,320	8,393	7,034	5,696	17,148
	Result	91.67	90.86	92.04	88.71	89.89	92.09	89.98	89.77
R2	#imgs	17,285	14,003	12,555	14,454	9,059	8,876	8,154	16,376
	Result	70.24	80.29	75.7	67.49	91.81	83.81	85.88	83.7

TABLE II: Statistics and results on UT-ZAP50K Dataset (#imgs denotes the number of images).

UT-ZAP50K	Closure	Toecap	HeelHeight	Gender	Material
#imgs(labeled)	48,637	36,874	29,969	50,000	47,260
#imgs(multi-label)	6,191	20,065	0	2,543	16,223
Result(Multi-task)	82.3				
	89.56	58.56	42.11	89.23	64.78
Result(Single-task)	-	85.8	83.1	-	74.5

model or tiled single-piece) product images are used. The dataset consists of 8 tasks including 4 length tasks and 4 design tasks and the statistics of the dataset are shown in Table I. The challenge is divided into two rounds: attribute classification on single-model product images and attribute classification on a combination of single-model and tiled single-piece product images. We adopt a multi-task single-label strategy to tackle this challenge, the results are shown in Table I. From the results of the two stages, we select the dataset with the superior performance as the basis for annotating the MEPAVE dataset. For the design task, we merge the results from four fine-grained sub-tasks. At last, we acquire 4 annotations of attribute slots: *Neckline*, *Sleeve Length*, *Skirt Length*, *Pant Length*. We choose this dataset as our source for distant supervision due to the high-quality annotations, its balanced distribution, and the shared domain with our target dataset.

UT-ZAP50K [45]. This is a dataset with 50,000 catalog shoe images from Zappos.com², which focuses on fine-grained attribute comparisons. We separately count all annotated data and the data with multiple labels for five attributes (*HeelHeight*, *Closure*, *Gender*, *Material*, and *Toecap*) in the UT-ZAP50K dataset. The data distribution and experimental results are shown in Table II. We conduct experiments on these 5 attribute comparison tasks in this dataset. Since this dataset consists of single-label task (*HeelHeight*) and multi-label tasks (*Closure*, *Gender*, *Material*, and *Toecap*), we first adopt the approach of multi-label and joint training on multiple tasks, and the accuracy is 82.3%. Then, we separately test the results of the joint-training model on five sub-tasks. The results are shown in line 2 of multi-task results in Table II. For tasks with poor performance *Toecap*, *HeelHeight*, and *Material*, we train single-task models for them. We choose 3 attributes with high-performance results *HeelHeight*, *Closure*, and *Toecap* in the UT-ZAP50K dataset to serve as annotations for the MEPAVE dataset. The reason we select this dataset for distant supervision is that the product images (shoes, boots) included in it closely match the types and styles found within the MEPAVE dataset.

²<https://www.zappos.com/>

TABLE III: Statistics and results on iFashion Dataset (#imgs denotes the number of images).

iFashion	Color	Material	Pattern
#imgs	395,778/8,934	489,395/8,741	264,953/8,925
#imgs(Multi-label)	188,943/4,811	76,231/1,586	11,335/269
Result(Single-label)	77.74	62.42	71.53
Result(Multi-label-Mix)	59.81	43.91	-
Result(Multi-label)	56.72	40.28	-

TABLE IV: Statistics and results on Craml Images.

Crawl Images	Upper Height
The number of images	20,000
Result	91.24

iFashion [46]. This is a large-scale multi-label dataset, which is constructed from over one million fashion images with a label space that includes 8 groups of 228 fine-grained attributes in total. We conduct experiments on 3 attribute classification tasks in this dataset, which are *Color*, *Material*, *Pattern*. All annotated data and the data with multiple labels for three attributes are counted separately. The processed data statistics and experimental results are shown in Table III. We train the single-task model on these three tasks using single-label or multi-label methods. Regarding the multi-label results, the difference between "Multi-label-Mix" and "Multi-label" is that the former trains the model using all single-label and multi-label data, while the latter only uses multi-label data. We choose attribute with superior result *Color* in the iFashion dataset to serve as annotations for the MEPAVE dataset. We select this dataset for distant supervision as its *Color* attribute distribution closely aligns with our target dataset.

Crawl Images. So far, the slot *Upper Height* has not been annotated. We crawl 20,000 images related to *Upper Height* from the Internet and train a model with single-label data. The result is shown in Table IV.

Overall, we annotate all 9 image-dependent attribute slots with the distant supervision approach. We extract the annotated attribute values covered in product descriptions in MEPAVE as value candidates in 16 text-dependent attribute slots. We perform post-processing on these value candidates. Specifically, we incorporate some synonyms in value candidates to narrow the candidate scope. If a slot is not mentioned in either product descriptions or images, its value will be annotated as "none". Based on the above steps, we construct a new multimodal attribute value prediction dataset, MAVP, which can facilitate our new task MuJo-SF. Our dataset consists of 79k product description-image instances. The statistics are shown in Table V. We randomly split all the instances into a training set with 64,468 instances, a validation set with 7,340 instances, and a testing set with 7,312 instances. Next, we will analyze the quality of the annotations obtained through the distant supervision method combined with multiple datasets.

3) *Quality of Distantly Supervised Annotations:* Here, we compare the distantly supervised annotations in the product image with the manual annotations in the product description to evaluate the accuracy of distantly supervised annotations. The relative accuracy is presented here, as only a subset of

TABLE V: Statistics of our dataset.

Category	#Product	#Instance	#Attribute slots	
			#Text-slot	#Image-slot
Clothes	12,240	34,984	10	3
Pants	2,832	8,004	10	2
Dresses	4,567	12,915	9	4
Shoes	9,022	21,005	5	5
Boots	713	2,212	7	4
Total	29,374	79,120	16	9

TABLE VI: The accuracy of distantly supervised annotations (T-S and V-S denote training set and validation set, respectively. H-Height and U-Height denote Heel Height and Upper Height, respectively.).

	Closure	Color	Sleeve Length	Toecap	Neckline
T-S	60.97	68.21	80.65	57.83	57.82
V-S	58.53	70.58	75.56	61.54	53.45
	H-Height	Pant Length	Skirt Length	U-Height	
T-S	52.09	70.11	65.28	96.39	
V-S	53.82	81.43	69.23	94.33	

annotated images is available for accuracy computation. The corresponding product descriptions for this subset of images include manually labeled attributes that match those present in the product images. The results are shown in Table VI. We can observe promising results in accuracy among the three attribute slots, *Upper Height*, *Sleeve Length*, and *Pant Length*, while the results for attribute slots *Toecap*, *Neckline*, and *Heel Height* are disappointing. We hypothesize that the lower accuracy might be attributed to two factors. First, the shoe images in the UT-ZAP50K dataset are captured from a single angle, while our dataset contains images taken from multiple angles. Second, the fine-grained collar predictions in the FashionAI Attribute dataset result in accuracy degradation when mapping the consolidated annotation results to our dataset. We leave the question of how to extract product attribute values more accurately from multi-angle images to future research.

C. MuJo-SF Framework

In this section, we develop a new training paradigm for our new task MuJo-SF based on our dataset MAVP, which distinguishes between text-dependent and image-dependent modules. Figure 4 shows the overview framework³.

1) *Input Representation:* First, we use BERT [40] to encode product descriptions, text-dependent slots, and text-dependent values, respectively. For product descriptions $D = \{w_1, w_2, \dots, w_k\}$, we concatenate $\{[CLS], w_1, w_2, \dots, w_k[SEP]\}$ and encode them into hidden representations \mathbf{H}_D , where $\mathbf{H}_D \in \mathbb{R}^{k \times d_t}$. For text-dependent slots and values, we encode each attribute along with its corresponding value candidates. Specifically, for $\forall s_j^T \in \mathcal{S}_{text}$, we feed $\{[CLS], s_j^T, [SEP]\}$ to another encoder BERT_{fixed} with fixed parameters, and obtain the hidden representation \mathbf{h}_{s_j} of the special token [CLS], where $\mathbf{h}_{s_j} \in \mathbb{R}^{d_t}$. We employ the same method to obtain hidden representations \mathbf{h}_v for each

³Note that our target is not to propose a sophisticated model. Rather, we focus on proposing a new task and presenting a model to benchmark our dataset.

value candidate corresponding to the s_j^T slot, where $v \in \mathcal{V}(s_j^T)$ and $\mathcal{V}(s_j^T)$ denotes the value candidate set of s_j^T . Among them, we assume that \mathbf{h}_{v_j} is the ground truth value representation corresponding to s_j^T . We aim to improve generalization and prediction accuracy, especially for sparse values. Moreover, we concatenate all text-dependent slots $\{s'_1, s'_2, \dots, s'_{|\mathcal{S}_{text}|}\}$ and encode them into hidden representation \mathbf{h}'_{s_T} , where $s'_j = \{[\text{SLOT}_j], s_j^T\}$, $1 \leq j \leq |\mathcal{S}_{text}|$, $[\text{SLOT}_j]$ denotes a special token. Then the representation $\mathbf{h}'_{s_j} \in \mathbb{R}^{d_t}$ at the position of $[\text{SLOT}_j]$ is used as the global representation of slot s_j^T .

And then, we resize the image to 224×224 pixels and obtain its visual representation from an advanced pre-trained model Swin-Transformer [43], the Swin-Transformer block consists of a shifted window-based multi-head self-attention (MSA) module, followed by a 2-layer MLP with GELU nonlinearity in between. The visual representation is defined as $\hat{\mathbf{U}} = \text{Swin-Transformer}(I)$, where $\hat{\mathbf{U}} \in \mathbb{R}^{d_a}$, and d_a is the dimension of visual representation.

2) *Cross-modality Interaction*: This module is shown in Figure 4. After receiving the product description representation \mathbf{H}_D and product image representation $\hat{\mathbf{U}}$, we apply a cross-attention mechanism to fulfill the cross-modality interaction between textual and visual representations. Specifically, we first feed \mathbf{H}_D and $\hat{\mathbf{U}}$ into the linear projection layer, respectively, and get \mathbf{H}'_D and $\hat{\mathbf{U}}'$, where $\mathbf{H}'_D \in \mathbb{R}^{k \times d_a}$ and $\hat{\mathbf{U}}' \in \mathbb{R}^{d_t}$. And then, we calculate attention scores between \mathbf{H}'_D and $\hat{\mathbf{U}}$, where \mathbf{H}'_D works as the query matrix, while $\hat{\mathbf{U}}$ works as the key and value matrix, and define the visual-aware textual representation $\tilde{\mathbf{H}}_D$ as follows:

$$\mathbf{Z}_D = \text{softmax} \left(\frac{\mathbf{H}'_D \hat{\mathbf{U}}^\top}{\sqrt{d_k}} \right) \hat{\mathbf{U}} \quad (4)$$

$$\tilde{\mathbf{H}}_D = \text{LN}(\mathbf{H}_D + \mathbf{Z}_D \mathbf{W}_D), \quad (5)$$

where LN denotes the layer normalization function [47], and $\mathbf{W}_D \in \mathbb{R}^{d_a \times d_t}$, $\tilde{\mathbf{H}}_D \in \mathbb{R}^{k \times d_t}$. Similarly, we can obtain the textual-aware visual representation $\tilde{\mathbf{U}} \in \mathbb{R}^{d_a}$.

3) *Text-Dependent Joint Slot Filling*: In this section, we aim to jointly fill all text-dependent slots \mathcal{S}_{text} with product descriptions. We first get the local and global representations \mathbf{h}_{s_j} , \mathbf{h}'_{s_j} of s_j^T , the value and ground truth value representations \mathbf{h}_v , \mathbf{h}_{v_j} , and the updated textual representation $\tilde{\mathbf{H}}_D$ from Section III-C1 and III-C2, respectively. And then, we apply the attention mechanism to fulfill the interaction between slot and textual information. Specifically, we calculate attention scores between \mathbf{h}_{s_j} and $\tilde{\mathbf{H}}_D$, and between \mathbf{h}'_{s_j} and $\tilde{\mathbf{H}}_D$, which can obtain two text-aware slot representations $\tilde{\mathbf{h}}_{s_j}$ and $\tilde{\mathbf{h}}'_{s_j}$. We fuse them to get the unified slot representation:

$$\tilde{\mathbf{h}}_{s_j}^T = \text{LN}(\tilde{\mathbf{h}}_{s_j} + \tilde{\mathbf{h}}'_{s_j}) \quad (6)$$

After that, following [48], we use L2 norm to acquire the distances between text slot representation $\tilde{\mathbf{h}}_{s_j}^T$ and candidate

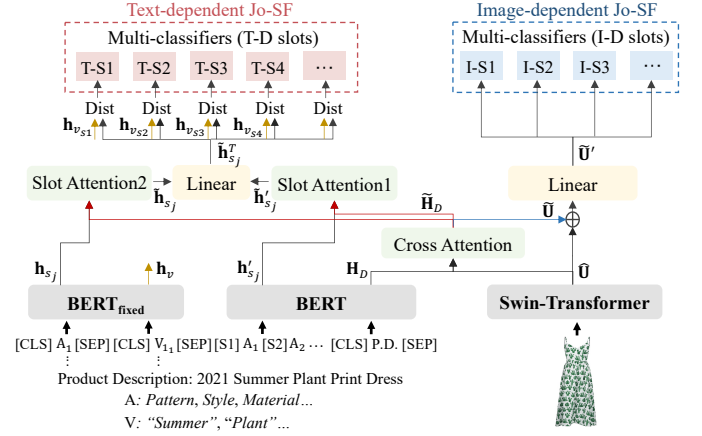


Fig. 4: Overview of our MuJo-SF framework. (T-D and I-D denote Text-dependent and Image-dependent. Dist denotes Distance, A and V denote Attribute and Value, and T-S1 and I-S1 denote the first slot in the text- and image-dependent, respectively).

values of s_j^T . The cross-entropy loss is taken as the training objective:

$$P_{slot_j}^T = \frac{\exp(-\|\tilde{\mathbf{h}}_{s_j}^T - \mathbf{h}_{v_j}\|_2)}{\sum_{v \in \mathcal{V}(s_j^T)} \exp(-\|\tilde{\mathbf{h}}_{s_j}^T - \mathbf{h}_v\|_2)} \quad (7)$$

$$\mathcal{L}^T = \frac{1}{|\mathcal{S}_{text}|} \sum_{j=1}^{|\mathcal{S}_{text}|} \text{CE}(P_{slot_j}^T, Y_{slot_j}^T) \quad (8)$$

where $P_{slot_j}^T \in \mathbb{R}^{|\mathcal{V}(s_j^T)|}$, $|\mathcal{V}(s_j^T)|$ denotes the number of values in the text slot s_j^T . $Y_{slot_j}^T$ is the ground truth label for the value corresponding to s_j^T , and $|\mathcal{S}_{text}|$ is the number of text-dependent slots.

4) *Image-Dependent Joint Slot Filling*: Here, we will jointly fill all image-dependent slots \mathcal{S}_{img} with visualized information in product images. We get the updated visual representation $\tilde{\mathbf{U}}$ from Section III-C2, which is concatenated with the original visual representation $\hat{\mathbf{U}}$ from Section III-C1. The result is then fed into the classifier corresponding to different image slots. Subsequently, we predict a value for each image slot. For image slot s_j^I , we predict the value corresponding to s_j^I :

$$\tilde{\mathbf{U}}' = \mathbf{W}_u [\tilde{\mathbf{U}}; \hat{\mathbf{U}}] \quad (9)$$

$$P_{slot_j}^I = \text{softmax}(\tilde{\mathbf{U}}' \mathbf{W}_j) \quad (10)$$

where $\mathbf{W}_u \in \mathbb{R}^{d_a \times 2d_a}$, $\mathbf{W}_j \in \mathbb{R}^{d_a \times |\mathcal{V}(s_j^I)|}$, $P_{slot_j}^I \in \mathbb{R}^{|\mathcal{V}(s_j^I)|}$, $|\mathcal{V}(s_j^I)|$ denotes the number of values in the image slot s_j^I . The cross-entropy loss is taken as the training objective:

$$\mathcal{L}^I = \frac{1}{|\mathcal{S}_{img}|} \sum_{j=1}^{|\mathcal{S}_{img}|} \text{CE}(P_{slot_j}^I, Y_{slot_j}^I) \quad (11)$$

where $Y_{slot_j}^I$ is the ground truth label for the value corresponding to s_j^I , and $|\mathcal{S}_{img}|$ is the number of image-dependent slots.

5) *Model Ensemble*: We employ a joint training method for the model. This approach is designed to simultaneously train both the text-dependent and image-dependent joint slot filling components. The overall objective function is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}^T + \lambda_2 \mathcal{L}^I \quad (12)$$

where λ_1 and λ_2 are hyper-parameters to control the contribution of each module.

IV. EXPERIMENTS

A. Experimental Setup

In this paper, we evaluate our proposed framework on our new dataset MAVP.

Evaluation Metrics. For our MuJo-SF task, we use Joint Goal Accuracy (JGA) and Slot Accuracy (SA) [48] as evaluation metrics. JGA assesses the model’s ability to accurately predict the entire set of attribute values for a product, with accuracy determined by the correct prediction of all slots. The JGA score is calculated as the product of indicator functions for all slots, and the formulation is shown below:

$$JGA = \prod_{j=1}^S I(y_j = \hat{y}_j) \quad (13)$$

SA serves as an effective metric for assessing the model’s performance on individual slots. The formulation of SA is shown below:

$$SA = \frac{1}{S} \sum_{j=1}^S I(y_j = \hat{y}_j) \quad (14)$$

Where S denotes the total number of slots to be tracked, for text-dependent slots, the total number of slots can be denoted as $|S_{text}|$, and for image-dependent slots, the total number of slots can be denoted as $|S_{img}|$. y_j is the true value for slot j , \hat{y}_j is the predicted value for slot j . The indicator function I returns 1 if the predicted value matches the true value, and 0 otherwise.

For these two metrics, JGA is stricter, which means that it only gives credit for completely correct outputs. This pushes models to be more precise and reduces the chance of overestimating performance due to partial correctness. SA provides a measure of how well the model performs on each individual slot. This can be useful for identifying which slots the model is handling well and which ones need improvement.

Implementation Details. We employ the pre-trained uncased BERT_{base} model [40] with dimension of 768 and the pre-trained Swin-Transformer [43] with dimension of 1024 to get the initial representations of text tokens and images, respectively. The image is resized to 224×224 pixels. For the joint training loss, we set the hyper-parameters $\lambda_1 = \lambda_2 = 1$ by tuning on the validation set. We employ an AdamW [49] optimizer for 50 epochs using a cosine decay learning rate scheduler and 20 epochs of linear warm-up. The initial learning rate and dropout rate are set to $2e-4$ and 0.3, which obtains the best performance on the validation set after conducting a grid search over the interval $[5e-5, 5e-4]$ and $[0.1, 0.6]$. The

TABLE VII: Results of our model compared with text and image classification baselines (T-D and I-D denote Text-dependent and Image-dependent, bk., L-M, BT-b, Ip-v4, and Sw-T denote backbone, LSTM, BERT-base, Inception-v4, and Swin-Transformer, respectively.). Our model achieves a statistically significant improvement with p-value<0.05 under a paired two-sided t-test.

Methods	Dev				Test			
	T-D slots		I-D slots		T-D slots		I-D slots	
	JGA	SA	JGA	SA	JGA	SA	JGA	SA
LSTM	86.04	99.29	-	-	86.49	99.30	-	-
BERT-base	86.97	99.31	-	-	87.21	99.32	-	-
Inception-v4	-	-	77.62	93.40	-	-	78.17	93.42
Swin-Transformer	-	-	80.16	93.98	-	-	80.53	93.99
Our model								
-bk.=L-M,Ip-v4	87.02	99.32	79.02	93.45	87.31	99.33	79.15	93.46
-bk.=BT-b,Ip-v4	87.89	99.35	79.11	93.46	88.10	99.36	79.22	93.47
-bk.=BT-b,CLIP	88.27	99.36	79.64	93.88	88.55	99.37	79.92	93.91
-bk.=BT-b,Sw-T	88.73	99.38	81.46	94.27	88.98	99.40	81.73	94.29

weight decay and warm-up are set to 0.02 and $5e-7$. Moreover, we set the batch size to 60 on one NVIDIA Tesla P40.

Baselines. Our model is compared with text baselines such as LSTM and BERT, and image classification baselines such as Inception-v4 [50], CLIP [51] and Swin-Transformer.

B. Main Results

Table VII shows the performance of our model compared with text and image classification baselines on Dev and Test sets. The results indicate that different baseline models yield promising performance on our new task, MuJo-SF, underscoring the effectiveness of the task and the high quality of the dataset. Moreover, integrating advanced pre-trained models significantly boosts the performance metrics. For instance, transitioning from LSTM to BERT-base for text-dependent slots, and from Inception-v4 to Swin-Transformer for image-dependent slots, resulted in noticeable gains. These results indicate that models enhance their ability to understand and process complex features by utilizing more sophisticated underlying architectures.

Meanwhile, we conduct experiments with another advanced encoder, CLIP; however, CLIP’s results on the JGA metric are slightly inferior to those of the Swin-Transformers. We guess this is because CLIP is designed to link visual content with textual descriptions and typically employs global self-attention across the entire image. In contrast, the Swin-Transformer focuses on traditional vision tasks and utilizes a combination of local and limited global self-attention mechanisms. Our task requires the extraction of information separately from text and image streams. Furthermore, product descriptions and images may contain varying attribute values for the same product, requiring a focus on various image regions for effective value extraction. Therefore, we consider the Swin-Transformer to be the more suitable model for our purposes. Future research could explore additional combinations of pre-trained models and delve into transformer-based architectures for multimodal learning.

And then, the consistently higher scores on the Slot Accuracy (SA) metric compared to the Joint Goal Accuracy (JGA) across all models emphasize the difficulty of achieving perfect predictions across all slots. While high SA scores indicate that models can accurately predict individual slots, the JGA metric reveals a lack of holistic understanding required to accurately predict a full set of attribute values. These results suggest that future research should focus on enhancing models' capabilities to capture and correlate comprehensive contextual information. Moreover, the challenge of achieving a high JGA score indicates potential improvements in error analysis. For example, future work could investigate whether errors are systematic or random, whether they relate to specific slots or types of data, or whether they are due to limitations in the models' representational power.

At last, our training mechanism achieves promising results across different dependent types, and our method of joint training on multiple dependent types outperforms their individual results (text-dependent or image-dependent). Future research could involve exploring the optimal balance between types of dependencies and extending these findings to other multimodal tasks.

C. Ablation Study

To show the effectiveness of each module in MuJo-SF, we conduct ablation study by removing particular component from it. Table VIII shows the results. We can observe that all components in our MuJo-SF contribute to the final results. First, after removing the cross-attention module, the performance drops on all metrics, particularly on the text-dependent type. Specifically, JGA scores on two dependent types degrade by 1.56% and 0.96%, respectively. The results indicate that the cross attention mechanism promotes effective interaction between image and text. As shown in Figure 3, based on the visualized saliency ascription in Section III-B1, we quantitatively estimate visualized information in images to fill image-dependent slots. However, many attributes such as *Material* and *Season*, whose visualized saliency scores fall below the threshold still provide valuable information to the text. In addition, we ablate I-D training (w/o I-D loss) or T-D training (w/o T-D loss). We observe that by adding only the image without training the image-dependent module, the model's performance on T-D slots slightly decreased, dropping from 88.98% to 88.35% in JGA. However, by adding only text without training the text-dependent module, there is a marginal decrease in T-D slots and the JGA scores only decrease by 0.42%. These results indicate that the visual information contained in the product images can assist in filling attribute slots more accurately for the text-dependent task, while the limited product attribute information in the textual descriptions can only provide very limited help for the image-dependent task.

D. Case Study

We conduct further analysis with two specific examples. As shown in Figure 5, in MAVE, the sequential labeling method is commonly employed. Here, images are typically

TABLE VIII: Ablation study of MuJo-SF on test set.

Methods	T-D slots		I-D slots	
	JGA	SA	JGA	SA
MuJo-SF	88.98	99.40	81.73	94.29
-w/o cross attention	87.42	99.34	80.77	94.00
-w/o T-D loss	-	-	81.31	94.27
-w/o I-D loss	88.35	99.37	-	-

TABLE IX: Performance of the Joint Slot Filling task on MEPAVE dataset.

Tasks	MEPAVE
	JGA
Jo-SF (text-only) (ours)	88.64
MuJo-SF (multimodal) (ours)	90.73

utilized as supplementary information to facilitate the model in extraction. In MuJo-SF, as shown in the green dotted box, fine-grained image analysis is performed as a preliminary step, followed by the presentation of a histogram displaying saliency ascription results for different attributes. Based on ascription results, image-dependent slots and text-dependent slots are distinguished. For example, in Figure 5 (a), conducting fine-grained analysis on the shoe image yields image-dependent slots: *Toecap*, *Closure*, *Upper Height*, *Color*, and *Heel Height*. Meanwhile, the product description provides two text-dependent slots: *Material* and *Style*. Sometimes, the information contained in product descriptions can be limited; for instance, both the descriptions include only two attributes each. Our MuJo-SF expands the product attribute values by directly obtaining more information from images. However, as depicted in Figure 5 (b), there's a challenge: the skirt has two colors, pink and white. Our experiments in Section III-B2 show that the performance of multi-label classification for colors is frustrating. As a result, we employ a single-label classification model, which predicts the skirt's color as white. We will solve this issue in future research, aiming for subsequent models to enhance the accuracy of filling multi-label attribute values.

E. Further Discussions

Effectiveness of the Novel Joint Slot Filling. To verify the effectiveness of the new task, Joint Slot Filling, proposed by us, we perform the Joint Slot Filling task on the MEPAVE dataset. The results are shown in Table IX. We can observe that our task yields promising results on the MEPAVE dataset, indicating the feasibility of the joint slot filling task⁴.

Accuracy of Joint Prediction on Image-Dependent Module. In the image-dependent joint slot filling module, we use Swin-Transformer as the backbone and conduct joint prediction on all image-dependent slots. To verify the accuracy of joint prediction on multiple slots, we predict each slot separately with the same backbone. The results are shown in Figure 6. In this experiment, we use the Slot Accuracy metric to evaluate

⁴We adapt the MEPAVE dataset to fit our task. Please note that due to the lack of annotations for images in MEPAVE, this experiment does not differentiate between text-dependent slots and image-dependent slots. All slots are derived from attributes annotated in product descriptions. The multimodal joint slot filling approach involves integrating images as auxiliary information with textual product descriptions. This experiment aims to validate the effectiveness of our proposed joint slot filling task.

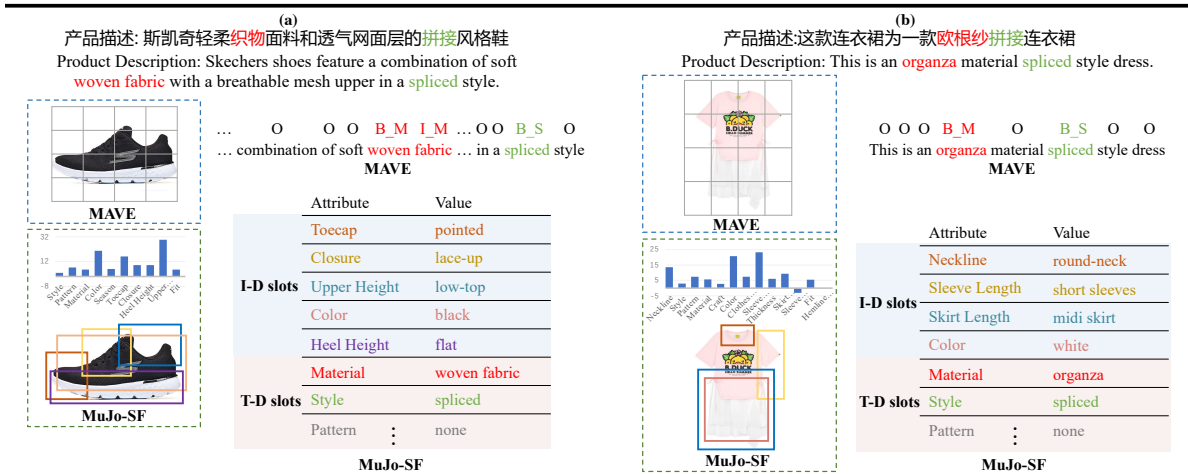


Fig. 5: Example comparison between MuJo-SF and MAVE (T-D and I-D denote Text-dependent and Image-dependent.)

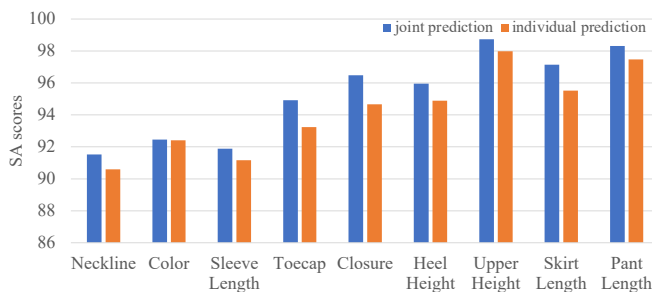


Fig. 6: Performance comparison of joint and individual predictions in Image-Dependent Module.

TABLE X: Performance comparison of different training mechanisms (OMFA denotes the method of one model fits all slots).

Methods	All slots	T-D slots	I-D slots
	JGA	JGA	JGA
OMFA	70.12	-	-
MuJo-SF	78.52	88.98	81.73

the performance. We can see that the joint prediction SA scores are better than the independent results in almost all slots. We guess it is because the joint prediction method can learn the dependencies among different slots. We choose not to present the results from the text-dependent joint slot filling module in this experiment due to the notably high SA scores, most of which exceed 99.1%. Such high scores result in minimal differences between joint and independent predictions. We speculate this may be due to the sparsity of the values in text-dependent slots.

Effectiveness of our New Training Paradigm. In this paper, we develop a new training mechanism for MuJo-SF, which distinguishes between text-dependent and image-dependent types. To verify the effectiveness of this training mechanism, we compare it with another training method we proposed (denoted as “one model fits all slots (OMFA)”), which unifies text and image representations into one representation with attention mechanism. The results are shown in Table X. Following

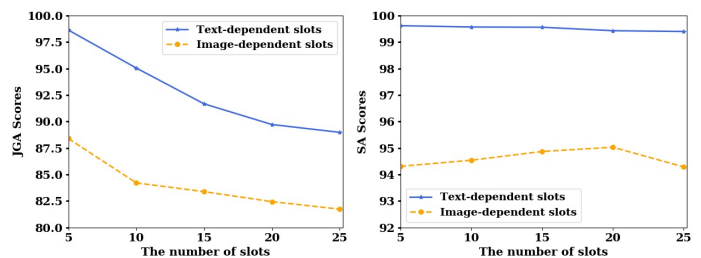


Fig. 7: Impact of different numbers of product attribute slots.

experimental analysis, the OMFA strategy exhibits an accuracy of 70.12% across all 25 slots in terms of the JGA metric. Subsequently, we conduct joint predictions for both text-dependent slots and image-dependent slots in MuJo-SF, requiring accurate predictions for both types of slots simultaneously. This approach yields a JGA result of 78.52%. Compared with our new training mechanism, the model’s performance is notably underwhelming when using the OMFA mechanism. There exists an 8.4% disparity between the results of OMFA and our training paradigm. We speculate that this discrepancy arises due to the rigor of the JGA metric, which demands the model to predict all slot values with absolute accuracy. Additionally, while the cross-attention mechanism is capable of integrating text and image information, the contributions from each source differ for various slots. This variation makes it difficult to utilize a single, unified representation to fill all slots at a high quality. This experiment indicates that how to design effective strategies for text-image fusion to accurately fill all text-dependent and image-dependent slots should be further explored in future research for the MuJo-SF task.

Impact of the Size of the Product Attribute Library. Here we will explore the impact of different numbers of product attribute slots on the model. We randomly select subsets of 5, 10, 15, and 20 attribute slots, ensuring a ratio of 60% text-dependent to 40% image-dependent slots, respectively (for example, among the 5 attribute slots, there are 3 text-dependent slots and 2 image-dependent slots), as well as the complete set of 25 attribute slots for this experiment. We use JGA and SA

metrics in this experiment. The results for text-dependent slots and image-dependent slots are illustrated in Figure 7. Given that different attribute slots include varying value candidates, we perform three sets of experiments, randomly selecting different subsets of slots for each. The experimental results are the average across these three sets.

We can find that the number of attribute slots does not significantly impact the results for the SA metric. This is because the SA metric assesses the model’s average performance across individual slots, and the model has already shown excellent results on this metric. For the JGA metric, the model achieves its highest result when only filling 5 attribute slots. As the number of attribute slots increases, the JGA results indeed tend to decline. However, as the number of slots increases, the reduction tends to level off. We hypothesize that this is because the JGA metric measures the model’s ability to accurately predict the entire set of attribute values for a product; as the number of attribute slots increases, predicting all slots correctly becomes increasingly challenging for the model. Our experiments indicate that for classification tasks, the scale of the classifier (as indicated by the number of attribute slots in our paper) has a discernible impact on the results measured by the stringent JGA metric.

In future work, we will explore developing a larger-scale multimodal product attribute value filling, enhancing the model’s ability to unseen product categories, and exploring the issue of missing attributes.

V. CONCLUSION AND FUTURE WORK

In this work, we propose a novel task MuJo-SF, aiming to combine the unique characteristics found in both product descriptions and product images to jointly fill values into pre-defined product attribute sets. This task presents two primary challenges: First, it necessitates simultaneous extraction and integration of text and image data, requiring sophisticated processing for accurate product attribute determination. Second, the strict Joint Goal Accuracy metric demands complete prediction across all slots, leaving no margin for error or bias toward any single attribute. The attribute values derived from our task can be applied to the construction of knowledge graphs, the refinement of recommendation systems, and the improvement of product retrieval in these real-world scenarios. To fulfill this new task, we develop a new dataset, MAVP, containing 79k product description-image instances and 25 slots across 5 domains. We present a strategy to fulfill visualized saliency ascription, which aims to distinguish between text-dependent and image-dependent attribute slots. For those image-dependent attribute slots, we annotate the corresponding attributes in images using distant supervisions. At last, we present a new training paradigm and use some baselines to benchmark the MAVP dataset. Our evaluation shows that the baseline models achieve promising results on two metrics. The new task leaves ample scope to promote future researches. For example, in future work, we can try to incorporate additional modules into the baseline model, such as an ensemble of image object detection tasks, to facilitate explicit alignment between fine-grained product image details and attribute slots.

REFERENCES

- [1] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, “Opentag: Open attribute value extraction from product profiles,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2018, pp. 1049–1058.
- [2] H. Xu, W. Wang, X. Mao, X. Jiang, and M. Lan, “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 (ACL)*, 2019, pp. 5214–5223.
- [3] J. Yan, N. Zalmout, Y. Liang, C. Grant, X. Ren, and X. L. Dong, “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 (ACL/IJCNLP)*, 2021, pp. 4694–4705.
- [4] G. Karamanolakis, J. Ma, and X. L. Dong, “Textract: Taxonomy-aware knowledge extraction for thousands of product categories,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 8489–8502.
- [5] Q. Wang, L. Yang, B. Kanagal, S. Sanghai, D. Sivakumar, B. Shu, Z. Yu, and J. Elsas, “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)*, 2020, pp. 47–55.
- [6] R. L. Logan IV, S. Humeau, and S. Singh, “Multimodal attribute extraction,” in *6th Workshop on Automated Knowledge Base Construction (AKBC@NIPS)*, 2017.
- [7] T. Zhu, Y. Wang, H. Li, Y. Wu, X. He, and B. Zhou, “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)*, 2020, pp. 2129–2139.
- [8] C. Bai, “E-commerce knowledge extraction via multi-modal machine reading comprehension,” in *International Conference on Database Systems for Advanced Applications (DASFAA)*, 2022, pp. 272–280.
- [9] R. Lin, X. He, J. Feng, N. Zalmout, Y. Liang, L. Xiong, and X. L. Dong, “Pam: Understanding product images in cross product category attribute extraction,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, 2021, pp. 3262–3270.
- [10] Q. Wang, L. Yang, J. Wang, J. Krishnan, B. Dai, S. Wang, Z. Xu, M. Khabsa, and H. Ma, “Smartave: Structured multimodal transformer for product attribute value extraction,” in *Findings of the Association for Computational Linguistics (EMNLP)*, 2022, pp. 263–276.
- [11] F. Li, Z. Wang, S. C. Hui, L. Liao, D. Song, J. Xu, G. He, and M. Jia, “Modularized interaction network for named entity recognition,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, 2021, pp. 200–209.
- [12] F. Li, Z. Wang, S. C. Hui, L. Liao, D. Song, and J. Xu, “Effective named entity recognition with boundary-aware bidirectional neural networks,” in *Proceedings of the Web Conference 2021 (WWW)*, 2021, pp. 1695–1703.
- [13] Y. Zheng, A. Hao, and A. T. Luu, “Jointprop: Joint semi-supervised learning for entity and relation extraction with heterogeneous graph-based propagation,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023, pp. 14541–14555.
- [14] Y. Zheng and A. T. Luu, “A novel, cognitively inspired, unified graph-based multi-task framework for information extraction,” *Cognitive Computation*, pp. 1–10, 2023.
- [15] E. F. Sang and J. Veenstra, “Representing text chunks,” in *9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1999, pp. 173–179.
- [16] Z. Huang, W. Xu, and K. Yu, “Bidirectional lstm-crf models for sequence tagging,” *arXiv preprint arXiv:1508.01991*, 2015.
- [17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” in *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, 2016, pp. 260–270.
- [18] J. P. Chiu and E. Nichols, “Named entity recognition with bidirectional lstm-cnns,” *Transactions of the association for computational linguistics*, vol. 4, pp. 357–370, 2016.
- [19] X. Ma and E. Hovy, “End-to-end sequence labeling via bi-directional lstm-cnns-crf,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

- [20] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A unified mrc framework for named entity recognition," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 5849–5859.
- [21] M. Jia, X. Shen, L. Shen, J. Pang, L. Liao, Y. Song, M. Chen, and X. He, "Query prior matters: A mrc framework for multimodal named entity recognition," in *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 3549–3558.
- [22] M. Jia, L. Shen, X. Shen, L. Liao, M. Chen, X. He, Z. Chen, and J. Li, "Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023, pp. 8032–8040.
- [23] X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, and J. Li, "Entity-relation extraction as multi-turn question answering," in *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, 2019, pp. 1340–1350.
- [24] S. Chen, Y. Wang, J. Liu, and Y. Wang, "Bidirectional machine reading comprehension for aspect sentiment triplet extraction," in *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 12 666–12 674.
- [25] L. Yang, Q. Wang, Z. Yu, A. Kulkarni, S. Sanghai, B. Shu, J. Elsas, and B. Kanagal, "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)*, 2022, pp. 1256–1265.
- [26] T. Nguyen, X. Wu, X. Dong, A. T. Luu, C.-D. Nguyen, Z. Hai, and L. Bing, "Gradient-boosted decision tree for listwise context model in multimodal review helpfulness prediction," in *Findings of the Association for Computational Linguistics (ACL)*, 2023, pp. 1670–1696.
- [27] C. Nguyen, T. Vu-Le, T. Nguyen, T. Quan, and A. T. Luu, "Expand BERT representation with visual information via grounded language learning with multimodal partial alignment," in *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 5665–5673.
- [28] T. Nguyen, X. Wu, A. T. Luu, Z. Hai, and L. Bing, "Adaptive contrastive learning on multimodal transformer for review helpfulness prediction," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 10 085–10 096.
- [29] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Transactions on Multimedia*, 2022.
- [30] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan, "Deepqoe: A multimodal learning framework for video quality of experience (qoe) prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3210–3223, 2020.
- [31] T. Nguyen, X. Wu, X. Dong, K. M. Le, Z. Hu, C.-D. Nguyen, S.-K. Ng, and A. T. Luu, "Read-pvla: Recurrent adapter with partial video-language alignment for parameter-efficient transfer learning in low-resource video-language modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024, pp. 18 824–18 832.
- [32] J. Wei, G. Hu, X. Yang, A. T. Luu, and Y. Dong, "Learning facial expression and body gesture visual information for video emotion recognition," *Expert Systems with Applications*, vol. 237, p. 121419, 2024.
- [33] H. Zhang, C. Yi, B. Zhu, H. Ren, and Q. Li, "Multimodal topic modeling by exploring characteristics of short text social media," *IEEE Transactions on Multimedia*, vol. 25, pp. 2430–2445, 2022.
- [34] L. Zhang, J. Shen, J. Zhang, J. Xu, Z. Li, Y. Yao, and L. Yu, "Multimodal marketing intent analysis for effective targeted advertising," *IEEE Transactions on Multimedia*, vol. 24, pp. 1830–1843, 2021.
- [35] Y. Han, L. Zhang, Q. Chen, Z. Chen, Z. Li, J. Yang, and Z. Cao, "Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 028–15 038.
- [36] X. Han, L. Yu, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang, "Fashionvil: Fashion-focused vision-and-language representation learning," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 634–651.
- [37] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 188–197.
- [38] Y. Chen, H. Zhong, X. He, Y. Peng, and L. Cheng, "Real20m: A large-scale e-commerce dataset for cross-domain retrieval," in *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 4939–4948.
- [39] J. Liu, Y. Chen, and J. Xu, "Saliency as evidence: Event detection with trigger saliency attribution," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 4573–4585.
- [40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018, pp. 4171–4186.
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.
- [42] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 6077–6086.
- [43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 10 012–10 022.
- [44] X. Zou, X. Kong, W. Wong, C. Wang, Y. Liu, and Y. Cao, "Fashionai: A hierarchical dataset for fashion understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (CVPR Workshops)*, 2019, pp. 296–304.
- [45] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014, pp. 192–199.
- [46] S. Guo, W. Huang, X. Zhang, P. Srikhanta, Y. Cui, Y. Li, H. Adam, M. R. Scott, and S. Belongie, "The imaterialist fashion attribute dataset," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2019, pp. 3113–3116.
- [47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [48] J. Xu, D. Song, C. Liu, S. C. Hui, F. Li, Q. Ju, X. He, and J. Xie, "Dialogue state distillation network with inter-slot contrastive learning for dialogue state tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023, pp. 13 834–13 842.
- [49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, 2014.
- [50] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017, pp. 4278–4284.
- [51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning (ICML)*, 2021, pp. 8748–8763.



Meihuizi Jia is a doctoral candidate at the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, and a visiting student at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Her research interests include natural language processing, multimodal information extraction and machine reading comprehension.



Jing Xu is a doctoral candidate at the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. He received his B.S. degree in information and computing science from the University of Mining & Technology, Beijing, China, in 2017. His current research interests include information extraction and knowledge graph.



Dr. Lei Shen is currently an algorithm engineer at JD.COM, China. Before that, she received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2022. Her research interests include natural language processing, dialogue systems, and the combination of large language models and recommendation systems. She has published over 20 papers in prestigious academic conferences such as AACL, ACM Multimedia, ACL, SIGIR, ICASSP, and CIKM. She also served as a Program Committee member for several top-tier

conferences including AACL, ACL, EMNLP, and ACM Multimedia.



Dr. Lejian Liao is currently a Professor with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. He received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1994. He has published over 40 papers in prestigious academic conferences such as ACL, AACL, IJCAI, ACM Multimedia, and NAACL. His research interests include machine learning, natural language processing, and intelligent networks.



Dr. Luu Anh Tuan is an Assistant Professor at NTU. Prior than that, he was a Research Fellow at MIT from 2018 to 2020. Luu's research interests lie in the intersection of AI, Deep Learning, and Natural Language Processing. He has published over 80 papers on top-tier conferences and journals including NeurIPS, ICML, ICLR, ACL, EMNLP, KDD, WWW, TACL, AACL, etc. He is the Associate Editor of Computational Linguistics journal and ACL Rolling Review. Luu also served as the Senior Area Chair of EMNLP 2020, Area Chair of ACL 2021-2024,

Area Chair of ICLR 2022-2023, Area Chair of NeurIPS 2023-2024, Senior Program Committee of IJCAI 2020-2021. He got the outstanding paper award in the International Conference on Learning Representations (ICLR) 2021. He was also a recipient of the Ministry of Trade and Industry (MTI) Singapore Innovation Award 2013.

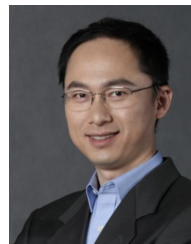


Shaozu Yuan is currently an algorithm engineer at JD.COM, China. His current research focuses on multimodal emotion analysis, natural language processing, and large language models. He has served as a reviewer for top-tier conferences including ACL, NeurIPS, and ACM MM.



Meng Chen is the AI Director of JD.COM. Before that, he was a Research Scientist of Nuance Communications. Meng Chen's research interests include natural language processing, speech recognition, and dialogue system. He has published over 40 papers in prestigious academic conferences such as AACL, IJCAI, ACM Multimedia, ACL, NAACL, ICASSP, Interspeech, and CIKM. He also serves as a Program Committee member for several top-tier conferences including AACL, ACL, EMNLP, NAACL, EACL, ICASSP, Interspeech, and ACM Multimedia. In 2023,

he was honoured with the Wuwenjun Artificial Intelligence Science and Technology Progress Award. He also received the Best Demo Award at ACM Multimedia in 2021.



Dr. Xiaodong He is the Vice President of JD.COM and Director of JD AI Research. He is also an Affiliate Professor at the University of Washington (Seattle), serves in doctoral supervisory committees. Before joining JD.COM, he was with Microsoft for about 15 years, served as Principal Researcher and Research Manager of the DLTC at Microsoft Research, Redmond, WA USA. His research interests are mainly in artificial intelligence areas including deep learning, natural language, computer vision, speech, information retrieval, and knowledge representation. He has published more than 200 papers in ACL, EMNLP, NAACL, CVPR, SIGIR, WWW, CIKM, NeurIPS, ICLR, ICASSP, Proc. IEEE, IEEE TASLP, IEEE SPM, and other venues. He received several awards including the Outstanding Paper Award at ACL 2015. He is a Fellow of IEEE. He received the bachelor degree from Tsinghua University (Beijing) in 1996, MS degree from Chinese Academy of Sciences (Beijing) in 1999, and the PhD degree from the University of Missouri-Columbia in 2003.