Cross-Modal Zero-Shot Product Attribute Value Generation

Jiaying Gong
Virginia Tech
Blacksburg, Virginia, USA
gjiaying@vt.edu

Ming Cheng
Virginia Tech
Blacksburg, Virginia, USA
ming98@vt.edu

Hongda Shen
eBay
New York City, USA
honshen@ebay.com

Pierre-Yves Vandenbussche
eBay
Amsterdam, Netherlands
pvandenbussche@ebay.com

Hoda Eldardiry
Virginia Tech
Blacksburg, Virginia, USA
hdardiry@vt.edu

ABSTRACT

Existing zero-shot product attribute value (aspect) extraction aims at using open-mining, graph, or large language models to predict unseen product attribute values. These approaches rely on uni-modal or multi-modal models, where the sellers should provide detailed textual inputs (product descriptions) for the products. However, manually providing (typing) the product descriptions is time-consuming and frustrating for the users. Thus, we propose a cross-modal zero-shot attribute value generation framework (ViOC-AG) based on CLIP, which only requires product images as the inputs. In other words, users only need to take photos of the products they want to sell to generate unseen attribute values. ViOC-AG follows a text-only training process, where a task-customized text decoder with a projection layer is trained with the frozen CLIP text encoder to alleviate the modality gap and task disconnection. During the zero-shot inference, product aspects are generated by the frozen CLIP image encoder connected with the trained task-customized text decoder. OCR tokens and outputs from a frozen prompt-based LLM correct the decoded outputs for out-of-domain attribute values. Extensive experiments with ablation studies conducted on the public dataset MAVE demonstrate that our proposed model significantly outperforms other fine-tuned vision-language models for zero-shot attribute value generation.

ACM Reference Format:
Jiaying Gong, Ming Cheng, Hongda Shen, Pierre-Yves Vandenbussche, and Hoda Eldardiry. 2024. Cross-Modal Zero-Shot Product Attribute Value Generation. In *Proceedings of the first workshop on Generative AI for E-Commerce 2024, October 25, 2024.* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Product attribute value (aspect) extraction aims at retrieving the values of attributes from the product's unstructured information (e.g. title, description), to serve better product search and recommendations for buyers. Existing uni-modal or multi-modal attribute value extraction models require users to manually provide (type) product descriptions, which is time-consuming and frustrating. In addition, these approaches mainly focus on supervised learning, weakly-supervised learning, and few-shot learning to train or fine-tune language models for attribute value prediction [12, 35, 36]. These approaches need labeled data for training and can not be extended to unseen attribute values for new products. To extract unseen attribute values, open-mining models [18, 35], inductive graph-based models [13], and multi-modal large language models [40, 41] try to generate potential attribute values from both product descriptions and images.

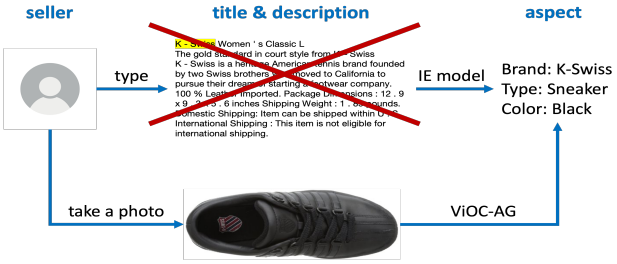

Figure 1: An example of cross-modal aspect generation.

However, these approaches suffer from these limitations: (1) It's difficult for classification or graph-based prediction models to scale to a large number of attribute values because the decision boundaries between classes become more complex and harder to learn, and increase the computational complexity. (2) Traditional information extraction models or the above multi-modal models need the inputs for product textual descriptions from the sellers (see Figure 1). It is challenging and time-consuming for the sellers to manually type and provide the product descriptions. To address the above limitations, we propose an optical character-enhanced zero-shot cross-modal model (ViOC-AG) to generate attribute values, which ONLY need the product images as the inputs. In other words, the seller only needs to take a photo of the product that he wants to sell without manually providing the product with textual descriptions, resulting in a better user experience.

There are two main challenges for zero-shot cross-modal aspect generation. The first challenge is the modality gap between vision and language caused by cross-modal generation. Although there exist many large generative image-to-text transformers (i.e. GIT [30], BLIP [17], BLIP-2 [16]), they target at the image captioning or visual question answering tasks. Our experiments in Sec. 4 show that simply fine-tuning these large vision language models performs poorly on the product attribute value generation task. This is because there is a task disconnection between language modeling (used for image captioning) and aspect generation. Thus, we take advantage of the pre-trained CLIP [25] ability to align visual and textual representations in a shared embedding space to avoid the modality gap. To alleviate task disconnection, we train a task-customized text decoder with a projection layer, which follows

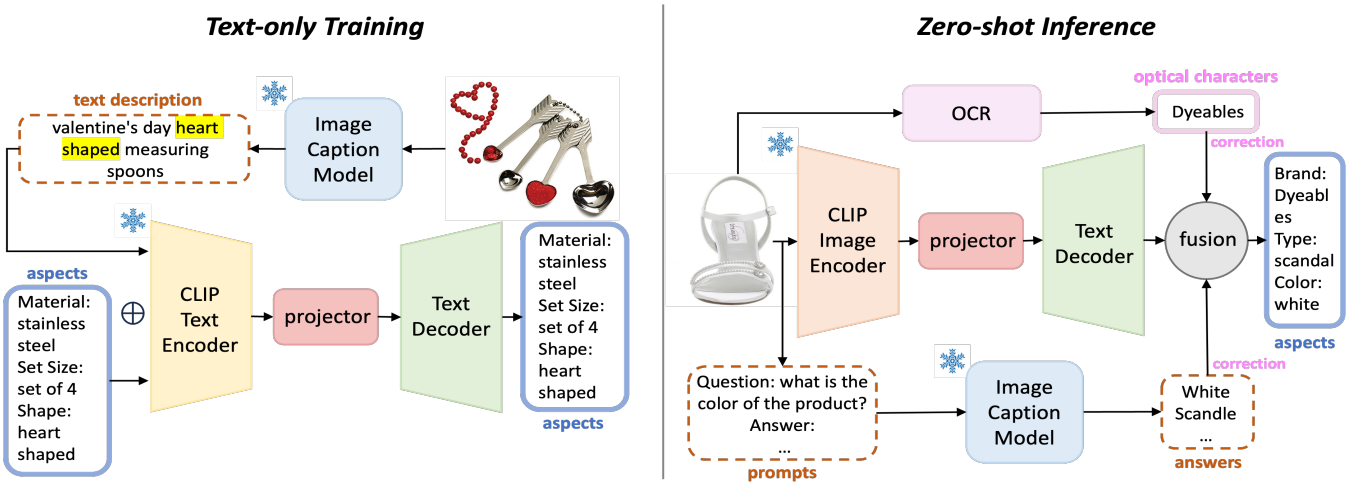

Figure 2: The overview of our proposed ViOC-AG model. Only the projector and the text decoder are trainable.

a text-only training process. Specifically, we tend to transfer CLIP textual description embeddings back into textual aspects by learning a task-customized decoder for the frozen CLIP text encoder using only text. The second challenge is the out-of-domain aspects caused by zero-shot generation. For zero-shot aspects, the model is susceptible to generate aspects that are not actually present in the input image but frequently appear during training (object hallucination). Due to the characteristics of the product attribute value generation task, some aspects (i.e. brand name, capacity, etc.) are shown directly on the product. Thus, we correct the generated outputs from the trained task-customized text decoder with the OCR tokens. For further final aspects correction, we generate potential attribute value answers by designing prompt templates for pre-trained visual question-answering LLMs. The effectiveness of each module is shown independently in Sec. 4.2. Extensive experimental results on a public dataset MAVE [37] show that our proposed model ViOC-AG significantly outperforms other existing cross-modal large language models (LLMs) for zero-shot attribute value generation. ViOC-AG achieves competitive results with generative LLMs with textual product description inputs, demonstrating the positive potential that users only need to take photos of the selling products for aspect generation.

2 RELATED WORKS

Existing works on product attribute value extraction mainly focus on supervised learning to train classification models [4, 6, 7], QA-based models [3, 21, 28, 32] or large language models [1, 2, 9]. Recently, some works use few-shot learning [12, 36] and weakly supervised learning [35, 39] to reduce the amount of labeled data for training. However, these approaches still need labeled data for multi-task training or iterative training. To extract unseen attribute values, open-mining models [18, 35] extract (explicit) attribute values directly from text, and zero-shot models [13] predict new attribute values by inductive link prediction of graphs. However, all these approaches can only extract attribute values from textual inputs. Some multi-modal models use both the product image and title with the description as the inputs to learn a better product representation for attribute value extraction or generation [11, 22, 23, 31, 33, 40, 41]. Though performance is improved by fusing more semantic information from multiple modalities, more input data is needed during the training stage. To enable image-first interactions from sellers and make it simple for the users, we propose a zero-shot cross-modal model motivated by image captioning [10, 14, 29, 34, 38] for attribute value generation, where only images are used as inputs.

3 METHODOLOGY

3.1 Problem Definition

Cross-modal attribute-value generation aims at automatically generating textual product attribute values from the product image. Consider a dataset $\mathcal{D} \subset \mathcal{I} \times \mathcal{T}$ where $\mathcal{I}$ is the image domain and $\mathcal{T}$ is the text domain, and (I_i, A_i) forms a corresponding image-aspect pair (i.e. $A_i \in \mathcal{T}$ is attribute values from product I_i). It can be formalized as a sequence generation problem given an input image I with a set of detected OCR tokens T, the model needs to infer the attribute values $A = [a_1, \cdots, a_N]$. The problem focuses on searching A by maximizing $p(A|I)$:

$$\log p(A|I) = \log \prod_N p(a_N|I, T, a_{1:n-1}) \tag{1}$$

where T is the set of OCR tokens detected from the product image I. The training process is typically accomplished in a supervised manner by training on manually annotated datasets and optimizing weights to converge to the optimal state. Therefore, it is necessary to explore optical-characters-aware zero-shot methods for guiding large-scale language models free of parameter optimization.

3.2 Zero-Shot Data Sampling and Pre-processing

For zero-shot attribute-value(aspect) generation, we follow [13] to let $A^S = [a_1^S, \cdots, a_N^S]$ and $A^U = [a_1^U, \cdots, a_N^U]$ denote the seen aspects and unseen aspects, where $A^S \cap A^U = \varnothing$. Because one product

may contain multiple aspects, We follow a generalized zero-shot setting [24] to ensure that any product in the validation/testing set has at least one aspect from A^U. For data pre-processing, we first combine the aspects that only have differences in uppercase/lowercase, singular/plural forms, or similar meanings and drop the data that we can not retrieve the corresponding images by the provided URLs in MAVE [37]. We implement the zero-shot data sampling over 21 categories of MAVE independently so that the zero-shot training, validation, and testing sets can still have similar data distributions across various categories.

3.3 Overall Framework

We introduce the overview of ViOC-AG in Figure 2, which is a transferable aspect generation framework based on CLIP [25] and trained on a text-only corpus. Specifically, we train a language decoder to decode the CLIP text embedding of aspects with generated text descriptions from a frozen image caption model. We make this decoding to be similar to the original textual aspects A. Namely, our training objective is a reconstruction of the input text from CLIP textual embedding. For zero-shot inference, we directly feed the CLIP image embedding of a given product image I into the trained decoder to generate aspects that are corrected by detected optical characters and values from the generated text description.

3.3.1 Text-only Training. Our goal is to train a transferable task-customized language decoder with a projector. During the training phase, we freeze all the parameters of the CLIP text encoder. We only train the projector from scratch and fine-tune the decoder-only language model (i.e. GPT-2) in predicting product attribute values. We first concatenate the generated descriptions of the product image via a frozen image caption model with the textual aspects inputs sequentially to prevent model overfitting and improve the generalization and robustness of the model. Next, we mapped the textual embeddings to CLIP space by CLIP text encoder E_T^*. Then, the projected text embedding is decoded back by a trainable decoder D_T. The text-only training objective is thus to minimize:

$$\sum_{A \in \mathcal{T}} \mathcal{L}(D_T(W \cdot E_T^*(A \oplus M^*(I)) + b), A) \tag{2}$$

where $*$ denotes a frozen model with parameters not updated during training. M^* can be any frozen image caption model (i.e. BLIP-2), and I is the product image. The projector $W(\cdot) + b$ is a learnable linear layer for domain alignment and dimension adjustment. $\mathcal{L}$ is an autoregressive cross-entropy loss for all tokens in A. The trainable projection layer alleviates the modality gap connecting the image domain with the text domain, and the task-customized text decoder solves task disconnection.

3.3.2 Zero-shot Inference. After the decoder D_T is trained, we leverage it for zero-shot aspect generation inference. Given a test product image I, we first extract its visual embeddings via the frozen CLIP image encoder E_I^*. We employ the trained projector and text decoder D_T to convert the visual embeddings into textual aspects:

$$A_D = D_T(W \cdot E_I^*(I) + b) \tag{3}$$

where $W(\cdot) + b$ is the trained projector for modality gap alleviation.

To improve the zero-shot performance caused by the out-of-domain attribute values, a fusion module is employed to correct the outputs from the text decoder D_T. We use information from two major sources to correct the outputs from A_D for the final aspects: (1) the values generated by the frozen prompt-instructed image caption model $A_P = \text{LLM}(I, P)$, where LLM can be any frozen cross-modal model (i.e. BLIP-2, InstructBLIP, etc.) [1], and P are the prompt templates (i.e. Question: What is the *attribute* of the product? Answer:). The *attribute* is replaced with the collected attribute names (i.e. type, brand, color, size, etc.) in the training set; (2) the optical characters T detected by the OCR module: [2]

$$T = \text{OCR}(I) = \{t | c_t > \tau_c\} \tag{4}$$

where c_t is token confidence value, and τ_c is the confidence threshold. In most cases, product attributes are from a known set (i.e. type, color, brand, capacity, etc.), only the values (i.e. long wallet, red, Chanel, 13oz, etc.) vary for different products and may include zero-shot cases, such as a new brand. We first check whether the attribute exists in the training set to decide whether the attribute is a zero-shot case or not. When the attribute is not a zero-shot case, we further compare the cosine similarity between A_D and A_P. If the value is closer to 1, A_P is used to correct A_D for irrelevant tokens. If they are quite different, we think it is a value zero-shot case, where OCR tokens T are used to further correct A_D. For attribute value zero-shot cases, only OCR tokens T are used to correct A_D because no relevant prompts are provided for the generated A_P. Details of the correction process are shown in Algorithm 1. The correction process solves the hallucination problem and improves the zero-shot performance on out-of-domain attribute values.

Algorithm 1: Zero-shot Inference Correction

Input : Aspects A_D, A_P, OCR tokens T and distance threshold τ_d
Output: Final Aspects A
for a_D *in* A_D **do**
  if *get_attribute*(a_D) $\in$ *get_attribute*(A_P) **then**
    if *cosine_similarity*(*get_value*(a_D), *get_value*(a_P)) $> \tau_d$ **then**
      A.update(a_P)
    else
      A.update($a_i | max(\text{cosine\_similarity}(a_D, a_P || T))$)
  else
    A.update($a_i | max(\text{cosine\_similarity}(a_D, T))$)
return A

4 EXPERIMENTS

4.1 Experimental Setup

We evaluate our model over a public dataset MAVE, which is a large e-commerce dataset derived from Amazon Review Dataset [37]. To simulate the zero-shot situation, we reconstruct the dataset into zero-shot learning settings followed by Sec. 3.2. The dataset statistics are shown in Table 1, where aspects are attribute values. We compare our model ViOC-AG with the following SoTA generative large language models: (1) text-only models BART [15] and T5 [27], (2) image-to-text models ViT-GPT [8, 26], GIT [30], LLaVA [20],

[1] We use BLIP-2 as the image caption model in our paper.
[2] https://github.com/JaidedAI/EasyOCR

Table 1: Dataset Statistics.

	Train	Validation	Test
Products	403005	94426	188267
Attributes	620	560	576
Aspects	44505	20148	33060

Table 2: Experimental results (%) of text-only models and image-to-text models on the MAVE dataset.

	80%Acc.	Macro-F1	Micro-F1	ROUGE1
BART	**79.32**	13.24	19.54	**60.59**
T5	68.69	**15.28**	**23.06**	53.82
ViT-GPT	16.60	2.62	4.07	31.00
GIT	14.89	3.70	5.36	34.13
LLaVa	25.67	7.20	10.24	40.11
BLIP	33.13	8.92	12.42	38.56
InstructBLIP	40.00	12.54	17.05	**44.20**
BLIP-2	45.85	13.92	18.86	43.06
ViOC-AG (ours)	**54.82**	**17.71**	**23.69**	31.92

BLIP [17], BLIP-2 [16] and InstructBLIP [5]. For evaluation, we use 80%Accuracy (we assume it is correct when 80% of the generated words are matched with the golden label for one aspect) to measure the generation accuracy. This is because the generative text decoder may generate more words than expected and we do not need a 100% accuracy rate, which means all generated tokens are exactly correct with the ground truth. Besides, we use Micro F1 and Macro F1 to evaluate the retrieval performance, which is a balance of Precision and Recall [3]. In addition, we use ROUGE1 [19] to evaluate the generation quality as ROUGE focuses on recall, which means how much the words in the ground truth appear in the candidate model outputs. Our model is implemented on PyTorch and optimized with AdamW optimizer. The learning rate is 0.0005. The batch size is 512. The cosine similarity threshold τ_d is 0.95, the OCR token confidence τ_c is 0.5. The experiments are conducted on eight Nvidia A100 GPUs with 80G GPU memory.

4.2 Results and Discussions

4.2.1 Main Results. The results of zero-shot attribute value prediction are shown in Table 2. We observe that: (1) In general, text-only models (BART and T5) show better performance than image-to-text models. This is because there is no modality gap for text-only models as they sacrifice the user experience that product text descriptions are needed for the model inputs. Thus, our goal is to build an image-to-text (cross-modal) model requiring only image inputs (product photos), which can achieve at least a similar performance to text-only models. (2) Although existing image-to-text LLMs (i.e. GIT, BLIP, LLaVa) have the zero-shot ability in image captioning, they perform poorly on product attribute value generation. We think that this is because there is a task disconnection between the image captioning task and the attribute value generation task. Simply fine-tuning the image-to-text LLMs may improve the image caption task. However, task-oriented information (i.e. task-customized decoder, OCR from the product, etc.) is also important for product attribute value generation tasks. (3) Our proposed model achieves the best Macro and Micro F1 scores among all text-only and image-to-text models, but it has a lower accuracy and ROUGE value compared with text-only models. We conjecture that this is because the trained task-customized text decoder may generate some non-relevant tokens, which reduces the percentage of the accurate tokens among all generated outputs, resulting in a low ROUGE and accuracy. More effective post-processing techniques can be studied in future work to remove the non-relevant tokens.

[3]We follow [40] to determine whether the generated answer is correct by checking whether the generated answer contains the true answer.

Table 3: Results (%) over ten categories on MAVE dataset.

	80%Acc.	Macro-F1	Micro-F1	ROUGE
Industrial	34.51	10.64	15.12	24.65
Home Kitchen	42.25	11.76	16.19	23.56
Automotive	43.64	13.28	17.49	28.81
Musical	51.74	14.65	20.08	30.76
Sports	47.38	16.08	21.73	30.16
Pet	64.45	20.62	28.51	36.44
Toys	61.19	23.25	30.54	41.75
Grocery	66.22	24.77	32.44	44.07
Clothing	63.63	25.14	33.30	42.58
Software	85.71	46.23	55.95	67.66

Table 4: Ablation results over ViOC-AG components in the zero-shot setting on MAVE dataset.

	80%Acc.	Macro-F1	Micro-F1	ROUGE
w/o D_T	38.34	12.23	16.71	22.47
w/o M^*	33.94	9.07	12.42	18.41
w/o prompts	49.63	15.71	21.07	27.36
w/o OCR	52.85	16.68	22.43	30.23
ViOC-AG (All)	**54.82**	**17.71**	**23.69**	**31.92**

We also conduct experiments across different categories of MAVE. Table 3 reports the selected categories (the worst 5 and best 5 categories). From Table 3, we observe that performance varies for different categories. Some categories (i.e. software, clothing, grocery) can achieve better performance because the products in these categories have optical characters shown on the surface of the product and different products have distinct patterns. Some categories (i.e. industrial, home kitchen, etc.) perform poorly because the patterns and features of the product images are quite similar and hard to distinct. For future work, a category-oriented training process can be explored to train category-related text decoders separately, where the OCR tokens, decoder outputs, and prompt answers can have different weights based on different categories.

4.2.2 Ablation Study. To verify the effectiveness of each component in ViOC-AG, we conduct the ablation study in Table 4. We observe: (1) The task-customized decoder and the frozen LLM used in the training phase play important roles in ViOC-AG as the performance drops drastically when removing these components. We

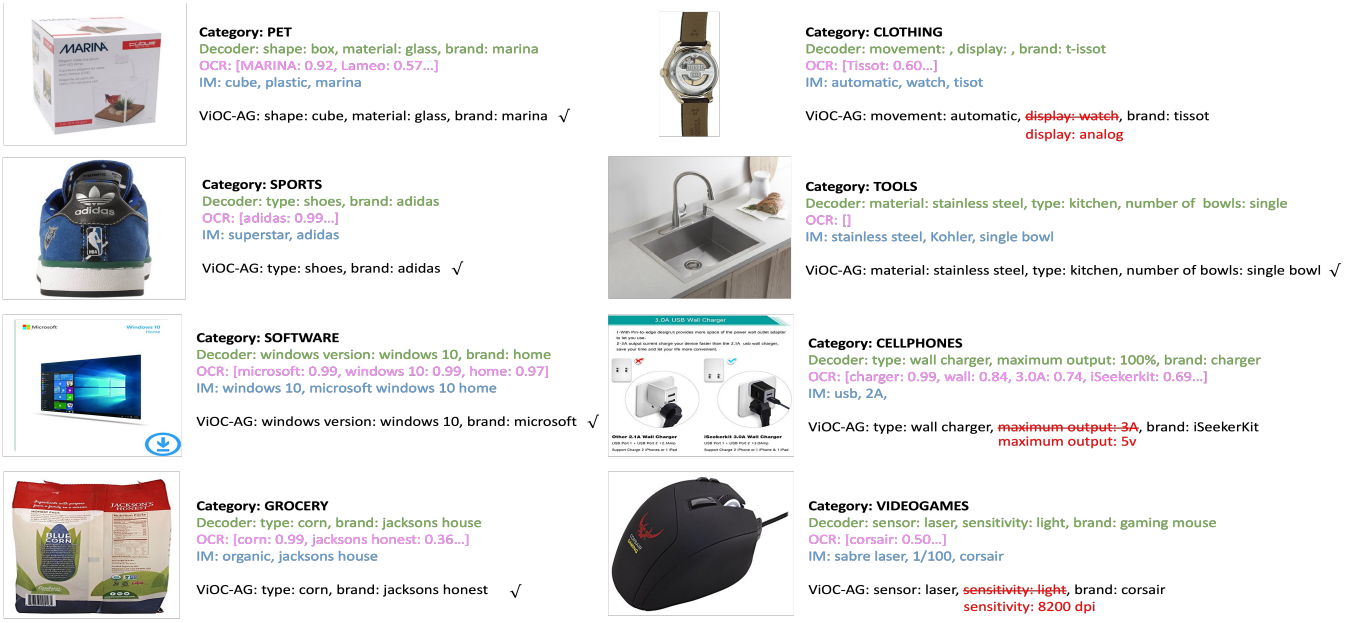

Figure 3: Demonstrations of ViOC-AG for product attribute value generation across eight different categories.

conjecture that it is because a pre-trained text decoder is usually used to generate long and diverse output text descriptions. However, our attribute value generation task is quite different where the generated outputs are short phrases with specific formats. In other words, there is no need for polishing the word (adding diversity) but correcting the phrase in the generation process. The outputs from the frozen LLM added to the original aspects inputs increase input data diversity, alleviating bias and overfitting for the trained text decoder. Thus, training a task-customized decoder with diverse inputs is essential. (2) Fusing answers from the frozen prompt-based LLM and OCR systems to correct the final generated aspects is useful for ViOC-AG, which is consistent with our hypothesis that some attribute values (i.e. brand name, capacity, etc.) may appear on the product packaging. To further improve the performance on out-of-domain aspect generation, a better customized OCR system, and diverse prompt templates can be explored in future work.

4.2.3 Case Study. Figure 3 shows the case study of our proposed model ViOC-AG on product attribute value generation across eight different categories. We observe that: (1) In general, most of the attribute values can be generated from the trained task-customized text decoder. There are some cases in which the trained decoder may not generate correct attribute values. For example, in the videogames case, the decoder generates 'gaming mouse' for the attribute of the brand. We conjecture that this is probably because of the data distribution and features of the training data. There are limited data (product) samples with the attribute value of 'brand: corsair' whereas there are lots of gaming mouse products in the training data. This issue is solved by our correction stage using OCR characters and answers from the image caption model introduced in Sec. 3.3.2. (2) OCR correction performs very differently among different categories. For the videogames case above, OCR can correct the brand name because 'corsair' is shown on the mouse. However, characters seldom appear for some categories such as TOOLS. In such categories, OCR shows limited or even no performance improvement. (3) In most cases, our proposed model ViOC-AG can correctly generate the attribute values after the correction stage for the trained text decoder. However, there still exists some difficult attributes such as 'display', 'maximum output', and 'sensitivity'. These attributes are never directly shown as characters in the image. In addition, these attributes can be hardly learned from the visual features of the product image. Such difficult cases have the following features: (a) Attribute names are rare in the training set. For instance, 'maximum output' and 'sensitivity' may only be applied to some specific products; (b) The values include digital numbers. If the digital numbers are not shown directly in the image, our OCR module can not help to correct the attribute values. The numbers (i.e. 5v, 8200 dpi, etc.) can not be learned from the visual features. These hard attributes need further exploration in future work.

5 CONCLUSION

We formulate the AVE as a cross-modal generation task, which only requires product images as the inputs. We propose an OCR-enhanced generation model ViOC-AG to generate unseen product aspects. ViOC-AG includes a text-only trainable projector and task-customized decoder to alleviate both the modality gap and task disconnection. For zero-shot inference, ViOC-AG employs OCR tokens and results from a frozen prompt-based LLM to correct the decoded outputs for out-of-domain attribute values. Experimental results on MAVE demonstrate that ViOC-AG outperforms other fine-tuned vision-language models and it can achieve competitive results with textual generative LLMs, showing the bright future directions of cross-modal zero-shot attribute value generation.

REFERENCES

[1] Nick Baumann, Alexander Brinkmann, and Christian Bizer. 2024. Using LLMs for the Extraction and Normalization of Product Attribute Values. *arXiv preprint arXiv:2403.02130* (2024).

[2] Alexander Brinkmann, Roee Shraga, Reng Chiz Der, and Christian Bizer. 2023. Product Information Extraction using ChatGPT. *arXiv preprint arXiv:2306.14921* (2023).

[3] Wei-Te Chen, Keiji Shinzato, Naoki Yoshinaga, and Yandi Xia. 2023. Does Named Entity Recognition Truly Not Scale Up to Real-world Product Attribute Extraction?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Mingxuan Wang and Imed Zitouni (Eds.). Association for Computational Linguistics, Singapore, 152–159. https://doi.org/10.18653/v1/2023.emnlp-industry.16

[4] Wei-Te Chen, Yandi Xia, and Keiji Shinzato. 2022. Extreme multi-label classification with label masking for product attribute value extraction. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. 134–140.

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems* 36 (2024).

[6] Zhongfen Deng, Wei-Te Chen, Lei Chen, and Philip S. Yu. 2022. AE-smnsMLC: Multi-Label Classification with Semantic Matching and Negative Label Sampling for Product Attribute Value Extraction. In *2022 IEEE International Conference on Big Data (Big Data)*. 1816–1821. https://doi.org/10.1109/BigData55660.2022.10020304

[7] Zhongfen Deng, Hao Peng, Tao Zhang, Shuaiqi Liu, Wenting Zhao, Yibo Wang, and Philip S. Yu. 2023. JPAVE: A Generation and Classification-based Model for Joint Product Attribute Prediction and Value Extraction. In *2023 IEEE International Conference on Big Data (BigData)*. 1087–1094. https://doi.org/10.1109/BigData59044.2023.10386204

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024. LLM-Ensemble: Optimal Large Language Model Ensemble Method for E-commerce Product Attribute Value Extraction. *arXiv preprint arXiv:2403.00863* (2024).

[10] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3136–3146.

[11] Pushpendu Ghosh, Nancy Wang, and Promod Yenigalla. 2023. D-Extract: Extracting dimensional attributes from product images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3641–3649.

[12] Jiaying Gong, Wei-Te Chen, and Hoda Eldardiry. 2023. Knowledge-Enhanced Multi-Label Few-Shot Product Attribute-Value Extraction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (<conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 3902–3907. https://doi.org/10.1145/3583780.3615142

[13] Jiaying Gong and Hoda Eldardiry. 2024. Multi-Label Zero-Shot Product Attribute-Value Extraction. *arXiv preprint arXiv:2402.08802* (2024).

[14] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10867–10877.

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[16] 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*. PMLR, 19730–19742.

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[18] Yanzeng Li, Bingcong Xue, Ruoyu Zhang, and Lei Zou. 2023. AtTGen: Attribute Tree Generation for Real-World Attribute Joint Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2139–2152.

[19] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[21] Hui Liu, Qingyu Yin, Zhengyang Wang, Chenwei Zhang, Haoming Jiang, Yifan Gao, Zheng Li, Xian Li, Chao Zhang, Bing Yin, et al. 2023. Knowledge-selective pretraining for attribute value extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 8062–8074.

[22] Mengyin Liu, Chao Zhu, Hongyu Gao, Weibo Gu, Hongfa Wang, Wei Liu, and Xucheng Yin. 2022. Boosting Multi-Modal E-commerce Attribute Value Extraction via Unified Learning Scheme and Dynamic Range Minimization. *arXiv preprint arXiv:2207.07278* (2022).

[23] Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. 2023. Multimodal pre-training with self-distillation for product understanding in e-commerce. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1039–1047.

[24] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. 2022. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence* 45, 4 (2022), 4051–4070.

[25] 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*. PMLR, 8748–8763.

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[27] 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 (2020), 1–67.

[28] Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. 2022. Simple and effective knowledge-driven query expansion for QA-based product attribute extraction. *arXiv preprint arXiv:2206.14264* (2022).

[29] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17918–17928.

[30] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100* (2022).

[31] Kai Wang, Jianzhi Shao, Tao Zhang, Qijin Chen, and Chengfu Huo. 2023. MP-KGAC: Multimodal Product Attribute Completion in E-commerce. In *Companion Proceedings of the ACM Web Conference 2023*. 336–340.

[32] 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*. 47–55.

[33] Qifan Wang, Li Yang, Jingang Wang, Jitin Krishnan, Bo Dai, Sinong Wang, Zenglin Xu, Madian Khabsa, and Hao Ma. 2022. Smartave: Structured multimodal transformer for product attribute value extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 263–276.

[34] Dongsheng Xu, Wenye Zhao, Yi Cai, and Qingbao Huang. 2023. Zero-TextCap: Zero-shot Framework for Text-based Image Captioning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4949–4957.

[35] Liyan Xu, Chenwei Zhang, Xian Li, Jingbo Shang, and Jinho D Choi. 2023. Towards open-world product attribute mining: A lightly-supervised approach. *arXiv preprint arXiv:2305.18350* (2023).

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

[37] 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*. 1256–1265.

[38] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23465–23476.

[39] Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jiawei Han. 2022. Oa-mine: Open-world attribute mining for e-commerce products with weak supervision. In *Proceedings of the ACM Web Conference 2022*. 3153–3161.

[40] 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* (2024).

[41] Henry Peng Zou, Gavin Heqing Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024. EIVEN: Efficient Implicit Attribute Value Extraction using Multimodal LLM. *arXiv preprint arXiv:2404.08886* (2024).