

EIVEN: Efficient Implicit Attribute Value Extraction using Multimodal LLM

Henry Peng Zou^{◇†}, Gavin Heqing Yu[♣], Ziwei Fan[♣], Dan Bu[♣],
Han Liu^{♡†} Peng Dai[♣], Dongmei Jia[♣], Cornelia Caragea[◇]

[♣]Amazon [♡]Washington University in St. Louis

[◇]University of Illinois Chicago
pzou3@uic.edu

Abstract

In e-commerce, accurately extracting product attribute values from multimodal data is crucial for improving user experience and operational efficiency of retailers. However, previous approaches to multimodal attribute value extraction often struggle with implicit attribute values embedded in images or text, rely heavily on extensive labeled data, and can easily confuse similar attribute values. To address these issues, we introduce EIVEN, a data- and parameter-efficient generative framework that pioneers the use of multimodal LLM for implicit attribute value extraction. EIVEN leverages the rich inherent knowledge of a pre-trained LLM and vision encoder to reduce reliance on labeled data. We also introduce a novel Learning-by-Comparison technique to reduce model confusion by enforcing attribute value comparison and difference identification. Additionally, we construct initial open-source datasets for multimodal implicit attribute value extraction. Our extensive experiments reveal that EIVEN significantly outperforms existing methods in extracting implicit attribute values while requiring less labeled data.

1 Introduction

Product attributes are crucial in e-commerce, aiding retailers in product representation, recommendation, and categorization, and assisting customers in product searching, comparison, and making informed purchasing decisions (Xu et al., 2019; Yan et al., 2021; Yang et al., 2023; Shinzato et al., 2023). Despite their importance, the accurate listing of these attributes remains a challenge. Sellers often fail to specify all relevant attribute values or list them incorrectly, leading to inefficiencies and potential customer dissatisfaction (Lin et al., 2021; Khandelwal et al., 2023). To address these issues, the task of Attribute Value Extraction (AVE) has

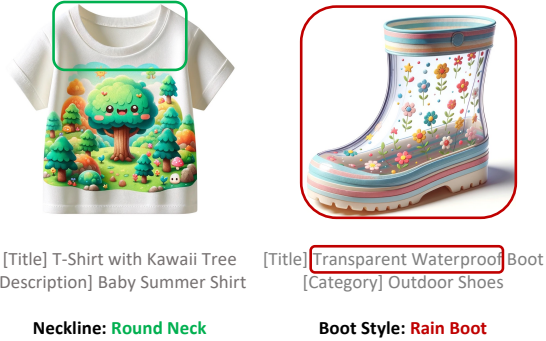


Figure 1: Examples of implicit attribute values. The attribute value cannot be explicitly extracted as a part of product texts, but can be inferred from the product image, text context or prior knowledge.

emerged as a key area of research in e-commerce. AVE seeks to automate the extraction of attribute values from product profiles such as product titles, descriptions, and images (Zheng et al., 2018; Wang et al., 2020, 2022).

Existing approaches for multimodal attribute value extraction can be broadly categorized into three categories: extractive, discriminative, and generative (more detailed discussion is provided in Appendix A). Most extractive studies focus on extracting attribute values that are explicitly stated in product text data (Zhu et al., 2020; Yang et al., 2022; Li et al., 2023; Xu et al., 2023). However, in real-world scenarios, an attribute value that needs to be obtained may not appear as a subsequence of the product text, but can be inferred from the product image, implied text context or prior knowledge about this product type (Zhang et al., 2023; Khandelwal et al., 2023; Blume et al., 2023). Take products in Figure 1 for example. The value “round neck” of the “neckline” attribute does not appear in product textual information, but can be easily identified from its product image. Similarly, the value “rain boot” corresponding to the attribute “boot style” in the second product is not explicitly stated but is implicitly embedded in its textual context

[†]Work done as an intern at Amazon.

“transparent waterproof” and visual information. In addition, previous discriminative and generative approaches for multimodal AVE are highly data-hungry, requiring large amounts of labeled data for training but still perform poorly in extracting implicit attribute values (Zhang et al., 2023; Fu et al., 2022). Furthermore, similar implicit attribute values are easily confused by the recent generative AVE model (Zhang et al., 2023).

To tackle these challenges, we introduce EIVEN, a data and parameter-efficient multimodal generative framework for multimodal implicit attribute value extraction. EIVEN utilizes the rich inherent knowledge of a pre-trained LLM and vision encoder to lessen reliance on extensive attribute-specific data. Additionally, to address the issue of model confusion caused by similar attribute values, we introduce a novel technique termed "Learning-by-Comparison". This approach feeds the model with pairs of instances that share the same attribute but potentially have different attribute values, forcing the model to compare and distinguish them.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first work to explore multimodal LLM for the emerging real-world problem of implicit attribute value extraction.
- We propose a novel Learning-by-Comparison technique to reduce model confusion among similar attribute values.
- We construct initial open-source datasets for multimodal implicit AVE.¹
- Extensive experiments show that our framework greatly outperforms recent multimodal AVE works, even with less labeled data.

2 EIVEN Framework

Problem Formulation. Given a product’s image and text context and a specified attribute, our goal is to extract the value for the corresponding attribute. Specifically, in our task of extracting implicit attribute values, the ground truth attribute value does not appear as a subsequence of the text context, but can be inferred from the product image, text context, or prior knowledge. In this work, we formulate the task of extracting implicit attribute values as the problem of generating answers given a question and product information. For example,

the question could be "What is the Sleeve Style of this product?" and the generated answer could be "Short Sleeve" by inferring from the product’s image and text context.

Figure 2 presents an overview of our efficient multimodal LLM, and Figure 3 illustrates our Learning-by-Comparison strategies. Next, we explain our key components in detail.

2.1 Image Embedding

We leverage projected multi-granularity visual features to serve as the visual token input to our LLM model. Specifically, we extract visual features from the $[cls]$ token in every M layer of the vision encoder and then concatenate them as:

$$I = \text{Concat}(\{I_k\}_{k=1}^K)$$

where K is the total number of extracted features, $I_k \in \mathbb{R}^{1 \times D}$ is the k -th extracted visual feature, and $I \in \mathbb{R}^{K \times D}$ is the overall multi-granularity image embedding.

Then, a simple visual projection network is used to adapt and transform the visual features to the same dimension as the text embedding of the LLM, which is denoted by:

$$I' = \sigma(IW_d + b_d)W_u + b_u$$

Here, $W_d \in \mathbb{R}^{D \times d_h}$ and $W_u \in \mathbb{R}^{d_h \times D_{text}}$ denote the weight matrices of the downsampling and upsampling layer, b_d and b_u are the bias terms, σ is the SwiGLU activation function (Shazeer, 2020; Luo et al., 2023b). In this way, we empower the LLM to understand visual features at multiple levels of granularity, such as edges, textures, patterns, parts, and objects (Ghiasi et al., 2022; Nguyen et al., 2019), which enables more effective extraction of attribute values.

2.2 Efficient Multimodal LLM

Previous generative works in multimodal implicit attribute value extraction (Zhang et al., 2023; Khandelwal et al., 2023) require large amounts of attribute-specific labeled data to achieve good performance. However, in the ever-evolving field of e-commerce, new products with unique attributes and values are constantly being introduced by different retailers and merchants. Gathering a large number of annotations for each new attribute is time-consuming and expensive (Yang et al., 2023; Lai et al., 2021; Zou and Caragea, 2023; Zou et al., 2023). To reduce reliance on labeled

¹<https://github.com/HenryPengZou/EIVEN>

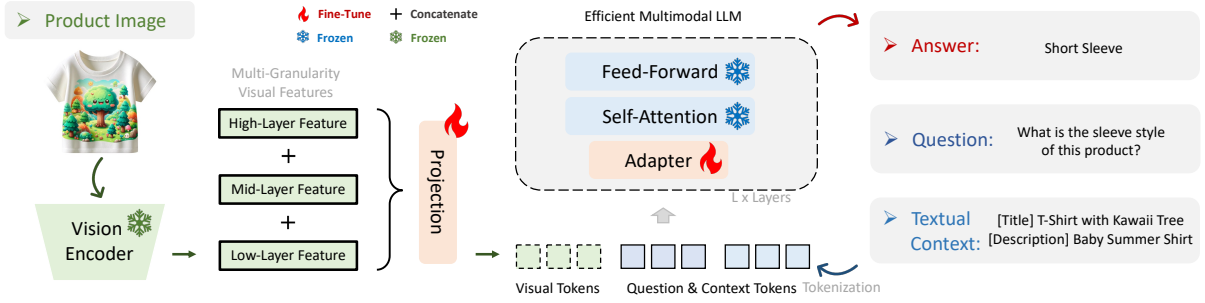


Figure 2: Overview of our efficient multimodal LLM. We extract multi-granularity visual features from a frozen pre-trained vision encoder and use a learnable visual projection network to align their dimensions with text token embeddings. The obtained visual tokens and tokenized question and text context are fed to the LLM (LLaMA-7B) to generate the answer. We insert lightweight adapters into every layer of the LLM for parameter-efficient fine-tuning.

Base Input	[Product Image] I; [Question] Q; [Text Context] C
Base	Q: What is the Sleeve Style of this product? A: Short Sleeve
LBC Input	[Product Image] I, I'; [Question] Q; [Text Context] C, C'
Judge_Last:	Q: What is the Sleeve Style of these two products? Are they the same? A: First: Short Sleeve ; Second: Long Sleeve ; No .
Judge_First:	Q: Do these two products have the same Sleeve Style ? Why? A: No ; First: Short Sleeve ; Second: Long Sleeve .
Better_Instance:	Q: Which product's Sleeve Style is Short Sleeve ? A: Second product has Short Sleeve .

Figure 3: Illustration of Learning-by-Comparison strategies. Our model is fed with pairs of product instances that share the same attribute but potentially different attribute values and asked to compare the values.

data, we pioneer the exploration of leveraging pre-trained LLMs for the multimodal implicit AVE task. Trained on vast and diverse datasets, LLMs have demonstrated remarkable understanding, generative capabilities, and few-shot transfer learning ability (Touvron et al., 2023; Liu et al., 2023; Wang et al., 2023; Tian et al., 2023; Dong et al., 2023; Lai et al., 2024), making them a promising approach to be explored for implicit attribute value extraction.

However, LLMs typically comprise billions of parameters, rendering their full-scale fine-tuning both resource-demanding and inefficient. To address this, we resort to parameter-efficient fine-tuning strategies, which has been proven to achieve performance comparable to full fine-tuning but with substantially fewer trainable parameters (Hu et al., 2023; Houlsby et al., 2019; Luo et al., 2023b; Tian et al., 2024). Specifically, we insert a lightweight adapter before every attention layer in our LLM. The mechanism of adapters is defined as:

$$h' = f_{\theta^u}(\sigma(f_{\theta^d}(h))) + h$$

where h, h' is the input and output of the adapter, $f_{\theta^d}(\cdot), f_{\theta^u}(\cdot)$ denotes for the downsampling and upsampling layers, σ is an optional activation func-

tion depending on the choice of adapters.

During training, we freeze all parameters in our LLM (LLaMA-7B (Touvron et al., 2023)) and the large image encoder, and only fine-tune these inserted lightweight adapters and the visual projection network.

Formally, given a product image embedding I , text context C , and an attribute-related question Q , the input of our multimodal LLM is denoted as $X = [I, Q, C]$. The overall training objective \mathcal{L} of our multimodal LLM can be defined as:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \sum_{t=1}^{|R|} \log p(R_t^i | X^i, R_{<t}^i; \theta_a, \theta_p)$$

where B is the batch size, R represents the ground-truth answer, R_t is the t -th token of R , $R_{<t}$ represents the tokens before R_t , θ_a denotes all parameters of adapters in LLM, and θ_p denotes all parameters in the visual projection network.

In our training scheme, although we use LLM, thanks to these lightweight adapters, the number of trainable parameters can be kept at a very small scale, e.g., 2~5M. This greatly reduces the memory requirement and allows efficient training of EIVEN on the same single 32G V100 GPU as the previous work (Zhang et al., 2023), while achieving significantly better performance even with much less labeled data.

2.3 Learning-by-Comparison

Many attributes have very similar attribute values, such as ‘Crew Neck’, ‘Scoop Neck’, and ‘Cowl Neck’, which can confuse models. To help models better distinguish these similar attribute values, we propose a new technique called Learning-by-Comparison (LBC) to assist model training.

Method	Approach	Clothing			Footwear			General			Average
		10	100	All	10	100	All	10	100	All	
M-JAVE (2020)*	Extractive	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CMA-CLIP (2022)	Discriminative	5.92	14.52	29.08	11.60	22.02	45.68	13.31	27.54	49.56	24.36
DEFLATE (2023)	Generative	13.29	25.23	56.52	11.43	35.94	74.80	9.75	39.22	59.11	36.14
EIVEN (Ours)	Generative	34.92	61.21	74.61	38.80	74.44	84.20	32.27	64.98	76.31	60.19
Absolute Gains (%p)	-	21.63	35.98	18.09	27.37	38.50	9.40	22.52	25.76	17.20	24.05

Table 1: Performance (micro-F1) comparison with representative work across different approaches. Models are trained with 10, 100, all (up to 1000) labeled data per attribute value. EIVEN delivers best results on all datasets, surpassing the latest implicit attribute value extraction work DEFLATE (Zhang et al., 2023) by 24.05%p on average. *Extractive approaches such as M-JAVE (Zhu et al., 2020) fail to handle implicit attribute values that do not appear explicitly as a subsequence of product text.

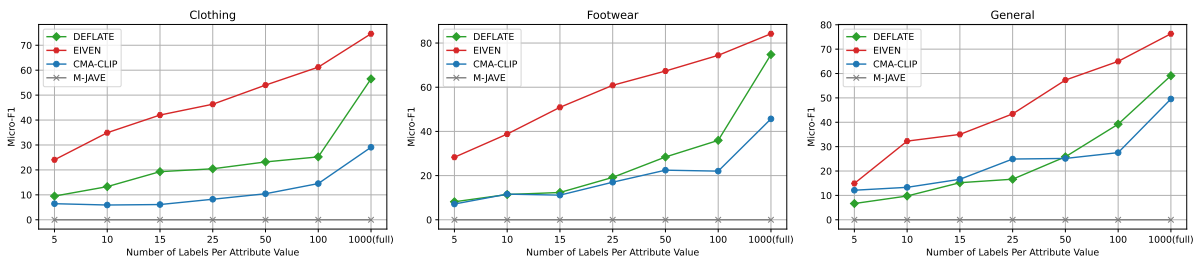


Figure 4: Data efficiency demonstration with varying numbers of labeled data. EIVEN can achieve better performance than DEFLATE with less labeled data, highlighting its data efficiency.

During training, in addition to the original product information I_1 , C_1 , and the query attribute A , we randomly sample another product with the same attribute A and include its image I_2 and text context C_2 in the model input for comparison. We have designed three strategies: LBC_Judge_Last, LBC_Judge_First, and LBC_Better_Instance as illustrated in Figure 3. We modify the attribute-related question and ground-truth answer accordingly. For example, in LBC_Judge_Last, we first ask the model to identify the value of the query attribute for both products, and then ask the model to compare and determine whether they have the same attribute value. The answer should be in the format of "First: {attribute value of the first product}; Second: {attribute value of the second product}; [comparison result]". Through this approach, the model is compelled to distinguish similar attribute values. Note that during the validation and testing phase, only the original product information and the attribute-related question are used.

3 Open-Source Multimodal Implicit AVE Dataset

Multimodal implicit AVE is an emerging problem, and there is currently a lack of truly open-sourced

datasets for multimodal implicit AVE.² Existing AVE datasets either do not contain product images or lack implicit attribute values. Thus, in this section, we introduce and make available several datasets to facilitate further research in this area.

Specifically, we present three multimodal implicit AVE datasets: Clothing, Footwear, and General. The statistics of these datasets are summarized in Table 6. All of them are derived and sampled from two publicly available datasets, MAVE (Yang et al., 2022) and Amazon Reviews 2018 (Ni et al., 2019). There are a total of 68,423 samples that cover 12 diverse product attributes and 87 common attribute values. Specifically, for each product attribute, we randomly collect product instances including the product texts (titles and product categories) and attribute values from the MAVE dataset. We collect popular attribute values with more than 100 instances for effective evaluation and randomly sample up to 1000 instances per attribute value to limit the dataset size. Since the MAVE dataset does not provide product images and is derived from the multimodal Amazon Reviews 2018 dataset, we collect the corresponding product images from the

²The claimed released multimodal implicit AVE dataset from DEFLATE (Zhang et al., 2023) is encrypted, and our multiple attempts to request decrypted data have failed.

Methods	MGVF	LBC	Image	Text	Clothing		Footwear		General		Average
					50	100	50	100	50	100	
EIVEN	✓	✓	✓	✓	54.01	61.21	67.33	74.44	57.31	64.98	63.21
- MGVF	✗	✓	✓	✓	49.92	57.75	65.04	72.73	53.5	62.27	60.20
EIVEN-Base	✗	✗	✓	✓	49.76	55.50	64.14	73.46	47.85	59.30	58.34
- Image	✗	✗	✗	✓	43.97	50.45	54.72	68.01	37.20	49.40	50.63
- Text Context	✗	✗	✓	✗	16.49	19.91	22.25	29.38	11.96	18.28	19.71

Table 2: Ablation study of key components and modality information. '50/100' represents the number of labels per attribute value, as is the case for the subsequent tables. "MGVF" denotes multi-granularity visual features.

Amazon Reviews 2018 dataset using their shared product identification number. Furthermore, the MAVE dataset contains only explicit attribute values. To evaluate performance on implicit attribute value extraction, we manually removed all explicit attribute value mentions from the product text for each product and its corresponding attribute. Therefore, attribute values in these data can only be inferred from product images, text context, or prior knowledge, i.e., implicit attribute values. Lastly, we split the train, test, and validation sets in a ratio of 0.75:0.15:0.15. We open-sourced these datasets.

4 Experiment

4.1 Experimental Setup

Baselines: We compare EIVEN with representative baselines in multimodal AVE: the latest generative work DEFLATE (Zhang et al., 2023), the representative discriminative work CMA-CLIP (Fu et al., 2022) and the extractive work M-JAVE (Zhu et al., 2020). Detailed descriptions of baselines are provided in Appendix C. **Metrics:** Following the latest work (Zhang et al., 2023), micro-F1 (%) is used as our evaluation metric and we determine whether the extraction results are correct using the exact match criteria, in which the full sequence of words is required to be correct.

Implementation Details: We select the ViT-B/16 (Dosovitskiy et al., 2021) of the pre-trained CLIP (Radford et al., 2021) as our image encoder. The multi-granularity visual features contain 4 [cls] tokens extracted from every 3 layer of ViT-B/16. We use LLaMA-7B (Touvron et al., 2023) as our LLM. The default dimension of the two-layer visual projection network is set to 128, and the dimension of the adapter in LLM is set to 8. LBC_Judge_Last is used as our default Learning-by-Comparison strategy. RepAdapter (Luo et al., 2023a,b) is adopted as our LLM adapter in default. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer and

Methods	Clothing		Footwear		General		Average
	50	100	50	100	50	100	
LBC_Judge_Last	54.01	61.21	67.33	74.44	57.31	64.98	63.21
LBC_Judge_First	53.08	60.25	66.78	74.64	54.97	64.71	62.41
LBC_Better_Instance	52.34	60.22	68.26	73.51	53.02	63.51	61.81
w/o LBC	49.76	55.50	64.14	73.46	47.85	59.30	58.34

Table 3: Ablation study on Learning-by-Comparison (LBC) strategies. All three strategies help improve performance, indicating their effectiveness in reducing model confusion. A visualization of the confusion matrix is provided in Appendix E.

train the model for 15 epochs. During the generation stage, we use top-p sampling as our decoding strategy with the temperature of 0.1 and the top-p value of 0.75. We report the micro-F1 result from a single run.

4.2 Performance Comparison with Baselines

The micro-F1 results with varying numbers of labeled data on the three multimodal datasets are shown in Table 1 and Figure 4. As can be seen from these comparison results, EIVEN can deliver significantly better performance on average than the other baseline methods. For instance, EIVEN can surpass the recent generative approach DEFLATE by 18.09%*p* on the Clothing dataset and 17.20%*p* on the General dataset. Also, EIVEN is much more data-efficient compared to previous generative attribute value extraction approaches. Using only 100 labels per attribute value, EIVEN can outperform or perform on par with other baselines trained with all labels (i.e., 1000 labels per attribute value) on all three datasets. These results indicate the effectiveness of our efficient multimodal LLM framework with the Learning-by-Comparison technique.

5 Ablation Study and Analysis

5.1 Effectiveness of Each Component

In order to quantify the impact of each component and modality in EIVEN, we measure and

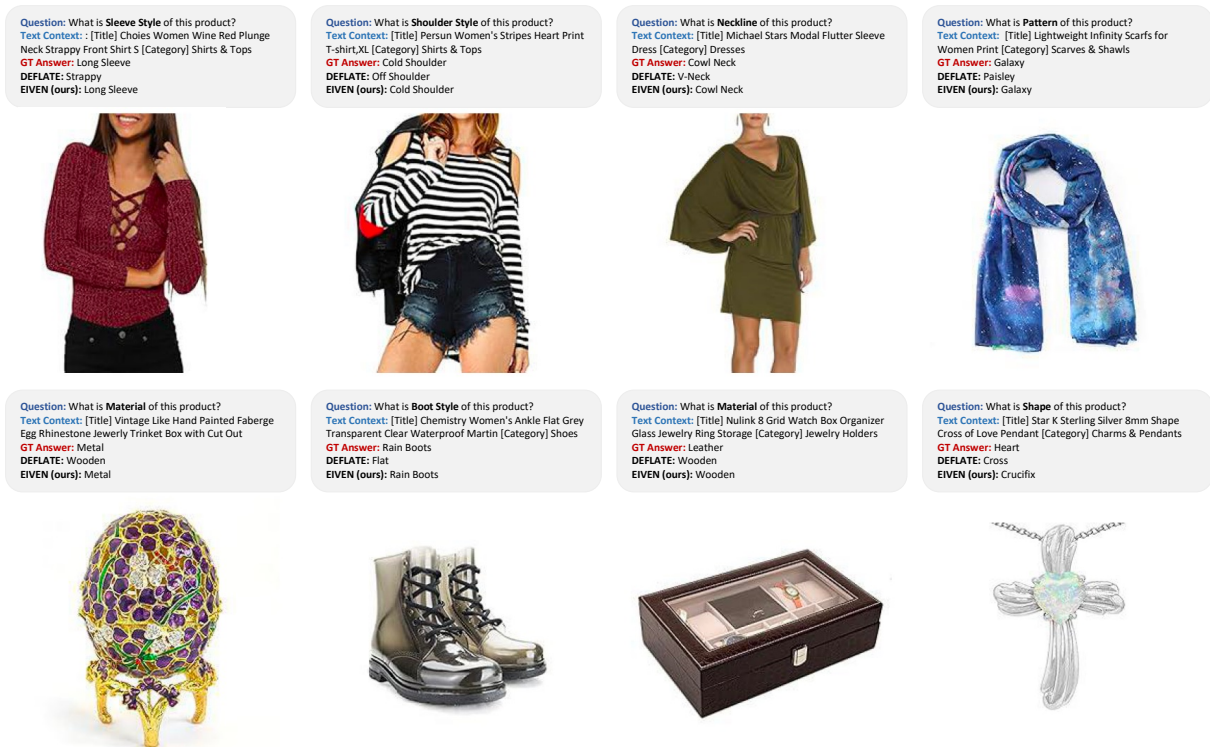


Figure 5: Qualitative examples and comparisons between EIVEN and DEFLATE.

summarize the micro-F1 result of EIVEN after removing different components and modalities in Table 2. First, we observe that the performance decreases after replacing multi-granularity visual features with the single-granularity feature or removing Learning-by-Comparison, suggesting that both of them contribute to the final performance of EIVEN. Notably, the performance of EIVEN-Base is still much better than DEFLATE, justifying the significant benefits of leveraging the LLM for implicit AVE. Besides, we can see that removing either the image or text context can significantly hurt model performance, which demonstrates the necessity of combining all these modalities in the implicit attribute value extraction task. Interestingly, the text modality plays the most important role, even when most of the ground truth attribute values cannot be explicitly identified from the product text. The possible reason is that implicit attribute values can still be inferred from the text context given the strong prior knowledge learned in LLM, as illustrated in the second product in Figure 1. On the other hand, extracting some product attribute values from images requires fine-grained visual understanding and thus is more challenging, especially when labels are limited.

5.2 Learning-by-Comparison Strategies

We explore different Learning-by-Comparison (LBC) strategies as illustrated in Figure 3. The results of these strategies are presented in Table 3. It is evident that all three strategies help improve the model's performance. This validates our motivation that including two instances into the model's input and asking the model to compare their attribute values can help alleviate model confusion among similar attribute values and improve overall performance. While there is no significant difference in performance among the three strategies, we believe that more effective LBC strategies can be devised to further enhance the model's performance, and we leave them for future exploration.

5.3 Qualitative Examples

Figure 5 demonstrates diverse qualitative examples and responses from the most recent generative work in implicit attribute value extraction DEFLATE and our method EIVEN. Compared to DEFLATE, EIVEN achieves overall better generation results across diverse product categories and attributes. In the first example, EIVEN extracts the correct attribute values for the product's sleeve style from the product image. In contrast, DEFLATE is confused by the strap in the neckline and generates

incorrect answers. In the sixth example, EIVEN demonstrates its ability to infer the correct value "Rain Boots" for the attribute "Boot Style" from the text context "Transparent Clear Waterproof Martin", prior knowledge, and product image. We also visualize some failure cases in the last two examples. We observe that EIVEN can make mistakes when multiple reasonable attribute values exist.

6 Conclusion

In this paper, we propose EIVEN, an efficient generative framework using multimodal LLM for implicit attribute value extraction. EIVEN leverages the rich internal knowledge of pre-trained LLM to reduce reliance on attribute-specific labeled data and adopts lightweight adapters for parameter-efficient fine-tuning of LLM. Besides, to enhance the visual understanding ability of our model, we feed multi-granularity visual features into LLM and propose Learning-by-Comparison strategies to alleviate model confusion among attribute values. We also release the first open-source dataset. Through extensive experiments on three multimodal implicit attribute value extraction datasets, we found that EIVEN can significantly outperform previous works using fewer labels, making it an efficient solution for implicit attribute value extraction.

Limitations

There are several limitations to our work. First, we only compared our approach with a limited number of baselines. This is because implicit multimodal attribute value extraction is a relatively new task, and also most of other multimodal attribute value extraction works are not open-sourced and very difficult to reproduce. We are planning to establish the first open-source benchmark for multimodal implicit AVE, which will also include comparisons among pre-trained general-purpose multimodal LLMs such as InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023) and GPT-4V. Second, we observed that some annotations from MAVE (Yang et al., 2022) are not accurate for implicit attribute value extraction, and there are some semantically overlapping attribute values. Automatic correction methods and human inspections are needed to construct more suitable benchmark datasets for implicit attribute value extraction. We plan to conduct such exploration in the future. In addition, more effective LBC strategies can be devised to further improve model performance.

References

- Ansel Blume, Nasser Zalmout, Heng Ji, and Xian Li. 2023. [Generative models for product attribute extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 575–585, Singapore. Association for Computational Linguistics.
- 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 (EC-NLP 5)*, pages 134–140, Dublin, Ireland. Association for Computational Linguistics.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tjong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [InstructBLIP: Towards general-purpose vision-language models with instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Zhikang Dong, Bin Chen, Xiulong Liu, Pawel Polak, and Peng Zhang. 2023. [Musechat: A conversational music recommendation system for videos](#). *ArXiv*, abs/2310.06282.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *International Conference on Learning Representations*.
- Jinmiao Fu, Shaoyuan Xu, Huidong Liu, Yang Liu, Ning Xie, Chien-Chih Wang, Jia Liu, Yi Sun, and Bryan Wang. 2022. [Cma-clip: Cross-modality attention clip for text-image classification](#). In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2846–2850.
- Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. 2022. [What do vision transformers learn? a visual exploration](#). *ArXiv*, abs/2212.06727.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for nlp](#). In *International Conference on Machine Learning*, pages 2790–2799. PMLR.
- Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. [LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

- Anant Khandelwal, Happy Mittal, Shreyas Kulkarni, and Deepak Gupta. 2023. [Large scale generative multimodal attribute extraction for E-commerce attributes](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 305–312, Toronto, Canada. Association for Computational Linguistics.
- Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng Cao. 2024. [Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2691–2701.
- Zhengfeng Lai, Chao Wang, Zin Hu, Brittany N. Dugger, Sen-Ching Samson Cheung, and Chen-Nee Chuah. 2021. [A semi-supervised learning for segmentation of gigapixel histopathology images from brain tissues](#). *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1920–1923.
- 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)*, pages 2139–2152, Toronto, Canada. Association for Computational Linguistics.
- Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. 2021. [Pam: Understanding product images in cross product category attribute extraction](#). In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3262–3270.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. 2023a. [Towards efficient visual adaption via structural reparameterization](#). *ArXiv*, abs/2302.08106.
- Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023b. [Cheap and quick: Efficient vision-language instruction tuning for large language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Anh Nguyen, Jason Yosinski, and Jeff Clune. 2019. [Understanding neural networks via feature visualization: A survey](#). *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 55–76.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. [Learning transferable visual models from natural language supervision](#). In *International conference on machine learning*, pages 8748–8763. PMLR.
- Noam M. Shazeer. 2020. [Glu variants improve transformer](#). *ArXiv*, abs/2002.05202.
- Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. 2023. [A unified generative approach to product attribute-value identification](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6599–6612, Toronto, Canada. Association for Computational Linguistics.
- Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and N. Chawla. 2024. [Tinyllm: Learning a small student from multiple large language models](#). *ArXiv*, abs/2402.04616.
- Yijun Tian, Huan Song, Zichen Wang, Haozhu Wang, Ziqing Hu, Fang Wang, N. Chawla, and Panpan Xu. 2023. [Graph neural prompting with large language models](#). In *AAAI Conference on Artificial Intelligence*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. [Learning to extract attribute value from product via question answering: A multi-task approach](#). In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 47–55.
- 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*, pages 263–276, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yu Wang, Nedim Lipka, Ryan A. Rossi, Alexa F. Siu, Ruiyi Zhang, and Tyler Derr. 2023. [Knowledge graph prompting for multi-document question answering](#). In *AAAI Conference on Artificial Intelligence*.

- Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. [Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223, Florence, Italy. Association for Computational Linguistics.
- Liyan Xu, Chenwei Zhang, Xian Li, Jingbo Shang, and Jinho D. Choi. 2023. [Towards open-world product attribute mining: A lightly-supervised approach](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12223–12239, Toronto, Canada. Association for Computational Linguistics.
- Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. [AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4694–4705, Online. Association for Computational Linguistics.
- Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, Sinong Wang, Zenglin Xu, and Dongfang Liu. 2023. [MixPAVE: Mix-prompt tuning for few-shot product attribute value extraction](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9978–9991, Toronto, Canada. Association for Computational Linguistics.
- Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. [Mave: A product dataset for multi-source attribute value extraction](#). In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 1256–1265.
- Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023. [Pay attention to implicit attribute values: A multi-modal generative framework for AVE task](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13139–13151, Toronto, Canada. Association for Computational Linguistics.
- Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. [Opentag: Open attribute value extraction from product profiles](#). In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1049–1058.
- Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. [Multimodal joint attribute prediction and value extraction for E-commerce product](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2129–2139, Online. Association for Computational Linguistics.
- Henry Zou and Cornelia Caragea. 2023. [JointMatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7290–7301, Singapore. Association for Computational Linguistics.
- Henry Zou, Yue Zhou, Weizhi Zhang, and Cornelia Caragea. 2023. [DeCrisisMB: Debiased semi-supervised learning for crisis tweet classification via memory bank](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6104–6115, Singapore. Association for Computational Linguistics.

A Detailed Discussion of Previous Works in Multimodal AVE

Existing approaches for multimodal attribute value extraction can be broadly categorized into three categories: extractive, discriminative, and generative (Table 4). **Extractive** approaches pose this task as a named entity recognition or sequence tagging problem, where the model outputs the start and end positions of the attribute value in the input text (Zhu et al., 2020; Xu et al., 2019). However, they are incapable of extracting implicit attribute values hidden in textual contexts or images. Additionally, they can only obtain raw value strings from product text, instead of the canonicalized values required for services such as faceted product search (e.g., 'Short Sleeve' instead of 'Short Sleeves' or 'Short Sleeved Shirt'). A further step is required for extractive approaches to canonicalize extracted raw value strings. **Discriminative** approaches classify each instance into a pre-defined set of attribute values (Fu et al., 2022; Chen et al., 2022). Yet, they cannot identify attribute values not in the pre-defined set and are hard to scale to large amounts of attributes. Ideally, we would like to eliminate the need to re-train a separate model for every new attribute or attribute value. **Generative** approaches frame the task as generating answers to attribute-related queries, using product information as a reference (Lin et al., 2021; Wang et al., 2022; Khan-delwal et al., 2023; Zhang et al., 2023). Given their nature of free-form text output, they are able to address implicit attribute values, unseen values, and can learn to directly obtain canonicalized values and answer values for multiple attributes. Nonetheless, previous generative methods in multimodal attribute value extraction require large amounts of labeled data for training and still perform very poorly on datasets with implicit attribute values.

Approach	Implicit Values	Unseen Values	Canonical Values	Scalable Attributes
Extractive	✗	✓	✗	✓
Discriminative	✓	✗	✓	✗
Generative	✓	✓	✓	✓

Table 4: Different AVE approaches and challenges.

B Dataset Statistics

The statistics of the introduced multimodal implicit AVE datasets (Footwear, Clothing, General) are provided in Table 6.

Methods	Linear	Sparse	# Param	Clothing		Footwear		Average
				50	100	50	100	
RepAdapter	✓	✓	1.70M	49.76	55.50	64.14	73.46	60.72
MLP-Adapter	✗	✗	2.23M	53.43	59.61	67.60	73.38	63.51
MLP-Adapter-L	✓	✗	2.23M	45.86	54.89	64.92	69.81	58.87

Table 5: Ablation study on the adapter in EIVEN-Base.

C Detailed Descriptions of Baselines

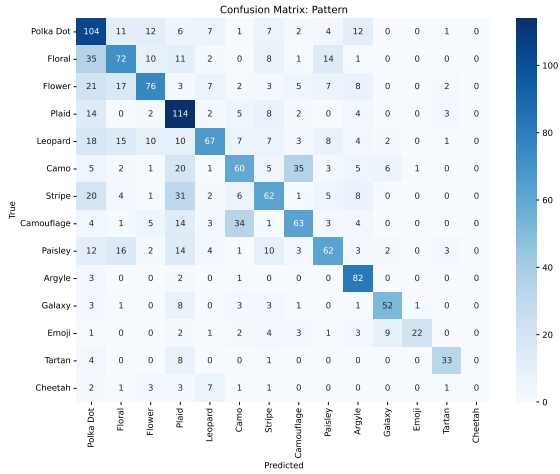
We describe in detail our baselines here: (1) **M-JAVE** (Zhu et al., 2020): A representative extractive approach that labels the input textual product description as "BIO" sequences related to attributes. It utilizes the fused multimodal features from the global and regional-gated cross-modality attention layer to make attribute predictions jointly. (2) **CMA-CLIP** (Fu et al., 2022): A recent discriminative approach that uses CLIP and sequence-wise attention to learn fine-grained multimodal product features. A modality-wise attention is then proposed to adaptively weigh the importance of visual and textual modalities to discriminate values for different product attributes. (3) **DEFLATE** (Zhang et al., 2023): A T5-based generative approach that consists of a generator to produce candidate attribute values from product information from different modalities and a discriminator to ensure the credibility of the generated answers.

D Ablation Study on Adapters

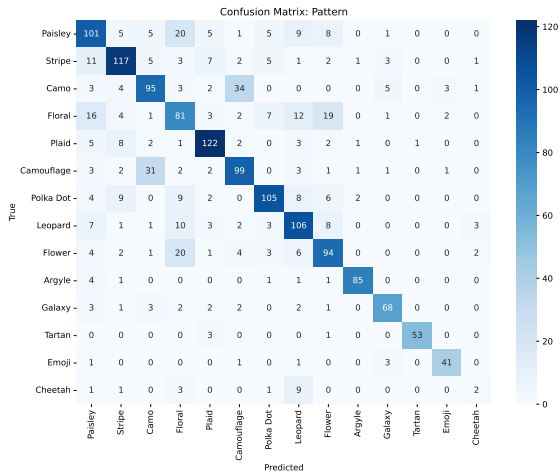
In this section, we study the performance of different types of adapters from the perspective of linearity and sparsity. RepAdapter (Luo et al., 2023a) is a recently proposed linear adapter without an activation function and has a sparse structure via group-wise transformation. The linear structure allows parameters in the adapter to be re-parameterized into LLM and thus introduces no inference latency. The sparse structure helps reduce the number of parameters and save memory consumption. Table 5 shows the comparison result with the representative MLP-adapter (Houlsby et al., 2019) in LLM. MLP-Adapter performs the best in micro-F1, while RepAdapter has the fewest parameters. We also observe that the linear structure generally sacrifices model micro-F1 performance in our task, and sparse transformation can boost model performance as well as reduce the number of parameters.

Dataset	# Samples	# Values	# Head	# Tail	Attributes
Footwear	26868	32	1000	229	Athletic Shoe Style, Boot Style, Shaft Height, Heel, Toe Style
Clothing	24664	30	1000	211	Neckline, Dress Length, Sleeve Style, Shoulder Style
General	16891	25	1000	117	Pattern, Material, Shape
Total	68423	87	1000	117	-

Table 6: Dataset statistics. ‘# Head’ and ‘# Tail’ denote the maximum and minimum amounts of attribute value instances among all attributes in the dataset. More details about these datasets can be found in Section 3.



(a) DEFLATE



(b) EIVEN

pared to DEFLATE, which validates our utilization of LLM and our Learning-by-Comparison strategy.

Figure 6: Confusion matrix for the Pattern attribute. LBC_Judge_Last is used in this example as the Learning-by-Comparison strategy. It can be observed that the confusion among attribute values is significantly reduced, demonstrating the effectiveness of our Learning-by-Comparison technique.

E Confusion Matrix

Figure 6 visualizes the confusion matrix of EIVEN and DEFLATE for the Pattern attribute on the General dataset using all labeled data. It can be observed that EIVEN has much less confusion com-