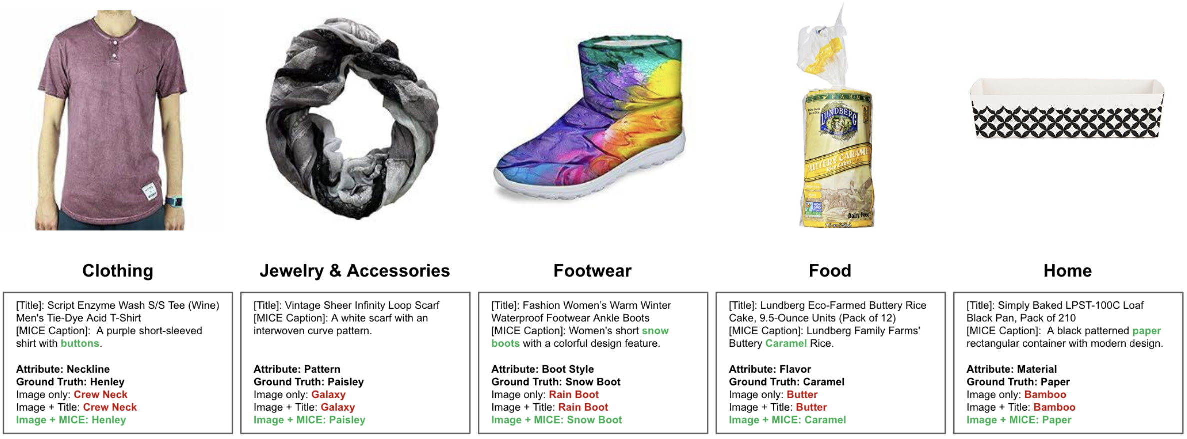
🐭🐭🐭MICE: Mixture of Image Captioning Experts Augmented e-Commerce Product Attribute Value Extraction

Jiaying Gong, Hongda Shen, Janet Jenq
eBay Inc.
`{jiagong,honshen,jjenq}@ebay.com`



[Title]: Script Enzyme Wash S/S Tee (Wine) Men's Tie-Dye Acid T-Shirt
[MICE Caption]: A purple short-sleeved shirt with buttons.

Attribute: Neckline
Ground Truth: Henley
Image only: **Crew Neck**
Image + Title: **Crew Neck**
Image + MICE: Henley

[Title]: Vintage Sheer Infinity Loop Scarf
[MICE Caption]: A white scarf with an interwoven curve pattern.

Attribute: Pattern
Ground Truth: Paisley
Image only: **Galaxy**
Image + Title: **Galaxy**
Image + MICE: Paisley

[Title]: Fashion Women's Warm Winter Waterproof Footwear Ankle Boots
[MICE Caption]: Women's short snow boots with a colorful design feature.

Attribute: Boot Style
Ground Truth: Snow Boot
Image only: **Rain Boot**
Image + Title: **Rain Boot**
Image + MICE: Snow Boot

[Title]: Lundberg Eco-Farmed Buttery Rice Cake, 9.5-Ounce Units (Pack of 12)
[MICE Caption]: Lundberg Family Farms' Buttery Caramel Rice.

Attribute: Flavor
Ground Truth: Caramel
Image only: **Butter**
Image + Title: **Butter**
Image + MICE: Caramel

[Title]: Simply Baked LPST-100C Loaf Black Pan, Pack of 210
[MICE Caption]: A black patterned paper rectangular container with modern design.

Attribute: Material
Ground Truth: Paper
Image only: **Bamboo**
Image + Title: **Bamboo**
Image + MICE: Paper

Figure 1: Examples of attribute value extraction for e-Commerce products.

Abstract

Attribute value extraction plays a crucial role in enhancing e-commerce search, filtering, and recommendation systems. However, prior visual attribute value extraction methods typically rely on both product images and textual information such as product descriptions and titles. In practice, text can be ambiguous, inaccurate, or unavailable, which can degrade model performance. We propose Mixture of Image Captioning Experts (MICE), a novel augmentation framework for product attribute value extraction. MICE leverages a curated pool of image captioning models to generate accurate captions from product images, resulting in robust attribute extraction solely from an image. Extensive experiments on the public *ImplicitAVE* dataset and a proprietary women's tops dataset demonstrate that MICE significantly improves the performance of state-of-the-art large multimodal models (LMMs) in both zero-shot and fine-tuning settings. An ablation study validates the contribution of each component in the framework. MICE's modular design offers scalability and adaptability, making it well-suited for diverse industrial applications with varying computational and latency requirements.

1 Introduction

Visual attribute value extraction is a fundamental task in e-commerce that involves identifying and structuring key, visually discernible product details, such as brand, size, color, material, and item specifications. Figure 1 shows a few examples of e-commerce products along with their images, titles and attribute key-value pairs from one public dataset. The attribute extraction process is critical for enhancing product visibility, improving search functionality, and enriching the overall consumer experience. Accurately extracted attributes improve search result relevance, boost product discoverability, and enable more precise product filtering, ultimately contributing to higher click-through rates and increased customer engagement.

In addition to improving search and discovery, structured product information helps consumers make more informed purchasing decisions by enabling easier comparison between similar items. On the back end, automated attribute extraction supports large-scale catalog management by mini-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1151–1160
July 28-30, 2025 ©2025 Association for Computational Linguistics

mizing the need for manual data entry, which can be inefficient and error-prone. It also promotes standardization across sellers and marketplaces, resulting in more consistent, high-quality product data while reducing operational overhead.

Existing research on attribute value extraction (AVE) has primarily focused on unimodal approaches, where product attributes are derived solely from textual inputs such as titles or descriptions (Gong and Eldardiry, 2024; Blume et al., 2023a; Gong et al., 2023; Shinzato et al., 2023; Yang et al., 2023a). More recently, multimodal methods leverage both product images and textual information with a joint learning framework to improve attribute extraction accuracy (Zou et al., 2024a; Liu et al., 2023b; Wang et al., 2023; Wu et al., 2023; De la Comble et al., 2022).

The reliance on seller-generated textual information presents challenges for e-commerce platforms. 1) Lack of standardization leads to inconsistencies in product formatting, making search and categorization difficult. 2) Incomplete attributes hinder filtering and recommendation systems, reducing product visibility. 3) Ambiguous or inaccurate descriptions contribute to misclassification and higher return rates, negatively impacting customer trust. These issues with seller-provided textual information negatively impact attribute value extraction performance (Chen et al., 2019; RetailTouchpoints, 2016).

Meanwhile, the rise of mobile listing apps on platforms like eBay, Amazon, and Alibaba has shifted seller behavior toward uploading images without structured text, streamlining the listing process through simplified and automated methods. In parallel, modern Large Language Models (LLMs) have demonstrated strong capabilities in generating high-quality titles and descriptions for e-Commerce listings (Zhang et al., 2024a,b; Chen et al., 2019). As a result, images are increasingly becoming the primary source of truth for attribute value extraction in e-commerce contexts.

In this work, we propose Mixture of Image Captioning Experts (MICE), an augmentation framework for attribute value extraction that leverages a mixture of image captioning models. Recent advances in Large Multimodal Models (LMMs) have proven effective at generating informative captions directly from images. Our hypothesis is that each independently trained LMM captures different visual aspects of a product, and their combined outputs can enrich the image signal with complementary information. These captions are then used to augment the input for attribute value extraction. Extensive experiments on the publicly available ImplicitAVE dataset show that our approach significantly improves performance across multiple state-of-the-art LMMs, outperforming models that rely solely on product titles. To assess the generalizability of this approach, we test MICE on an internal e-Commerce dataset, validating its effectiveness in real-world scenarios. Notably, our approach relies only on seller-provided images, yet achieves performance comparable to a proprietary closed-sourced LMM (i.e., GPT-4V) that ingests both images and product titles. This result highlights the potential of MICE as a vision-only alternative. Finally, ablation studies confirm the effectiveness of each individual component, and case studies illustrate how MICE produces accurate attribute values, even outperforming multimodal baselines.

2 Related Work

Most existing studies focus on extracting attribute values from product titles or descriptions by using classification models (Gong et al., 2023; Deng et al., 2022b,a), QA-based models (Liu et al., 2023a; Shinzato et al., 2022; Wang et al., 2020), transformers (Chen et al., 2023), hypergraphs (Hu et al., 2025a; Gong and Eldardiry, 2024), and generative LLMs (Gong et al., 2025a; Levine et al., 2024; Sabeh et al., 2024; Roy et al., 2024; Fang et al., 2024; Khandelwal et al., 2023; Shinzato et al., 2023; Blume et al., 2023b).

While prior models for product attribute value extraction primarily rely on a single modality, they often fail to capture the rich visual information and cross-modal correlations available in product images. Recent research has shifted toward leveraging Large Multimodal Models (LMMs), which jointly utilize product images and textual information to learn enhanced product representations for the AVE task. For example, product visual features are used to enhance product AVE by utilizing multi-modal transformers (Wang et al., 2022; Khandelwal et al., 2023), optical character recognition (Lin et al., 2021), multi-modal attention mechanisms (Zhang et al., 2023a; De la Comble et al., 2022), prompt-tuning of pre-trained transformers (Yang et al., 2023b), and LMMs (Hu et al., 2025b; Gong et al., 2025b; Zou et al., 2024b) that generate product attribute values from combined text and image inputs. To support an image-based

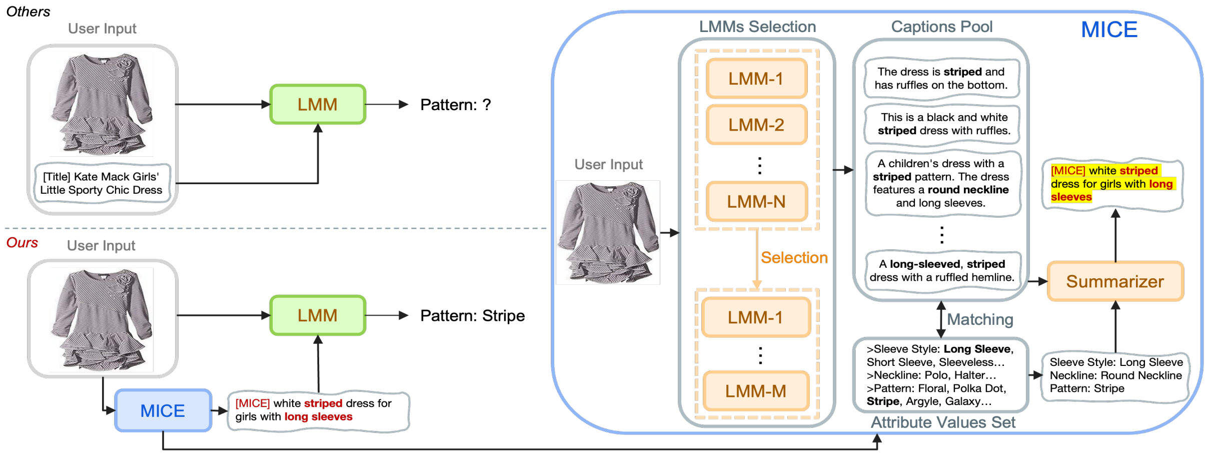

Others
User Input
[Title] Kate Mack Girls' Little Sporty Chic Dress
LMM
Pattern: ?
Ours
User Input
LMM
Pattern: Stripe
MICE
[MICE] white striped dress for girls with long sleeves
MICE
LMMs Selection
Captions Pool
User Input
LMM-1
LMM-2
LMM-N
Selection
LMM-1
LMM-M
The dress is striped and has ruffles on the bottom.
This is a black and white striped dress with ruffles.
A children's dress with a striped pattern. The dress features a round neckline and long sleeves.
A long-sleeved, striped dress with a ruffled hemline.
Matching
>Sleeve Style: Long Sleeve, Short Sleeve, Sleeveless…
>Neckline: Polo, Halter…
>Pattern: Floral, Polka Dot, Stripe, Argyle, Galaxy…
Attribute Values Set
Sleeve Style: Long Sleeve
Neckline: Round Neckline
Pattern: Stripe
Summarizer
[MICE] white striped dress for girls with long sleeves

Figure 2: The overview of MICE augmentation framework.

experience for sellers, our approach leverages the image captioning capabilities of modern LMMs. These generated captions expose implicit product information, effectively enhancing attribute value extraction especially when textual inputs are missing or unreliable.

3 Methodology

3.1 Problem Formulation

We consider the task of multimodal product attribute value extraction, where the input consists of a set of product/listing images $\mathcal{I} = \{I_1, \cdots, I_p : p \in \mathcal{P}\}$ and optional text inputs including descriptions and titles $\mathcal{T} = \{T_1, \cdots, T_p : p \in \mathcal{P}\}$ for each product $p \in \mathcal{P}$. The objective is to predict a value V_p for each target attribute A_p drawn from the attribute set $\mathcal{A} = A_1, \cdots, A_m$, where m denotes the total number of attributes (e.g., pattern, material, shape, etc). For a given attribute A_p, we define $\mathcal{L}_p$ as the set of possible candidate values.

3.2 Mixture of Image Captioning Experts

Mixture of Image Captioning Experts (MICE) leverages a pool of independently trained large multimodal models (LMMs) to generate high-quality captions for images. These captions are used to enrich the image-only modality and enhance attribute value extraction (AVE). An overview of MICE is given in Figure 2 highlighting its key components and demonstrating how it augments conventional LMM-based AVE approaches. MICE consists of three major components: (1) expert model selection, (2) caption generation and value matching, and (3) summarization of matched captions.

We first construct a pool of LMMs, denoted as $\mathcal{M} = M_1, \cdots, M_k$. For each input image I_p, every LMM in the pool generates candidate captions relevant to the attributes of interest. To retain only effective models, we evaluate each M_i on a held-out validation set using a predefined performance metric (e.g., micro-F1). We then filter out underperforming models using a threshold τ, resulting in the selected model set $\mathcal{M}_s$:

$$\mathcal{M}_s = \{\mathcal{M}_i | metric(\mathcal{M}_i) > \tau, \mathcal{M}_i \in \mathcal{M}\} \quad (1)$$

where $metric()$ is a pre-defined attribute value extraction performance metric e.g. micro-F1 and τ is the performance threshold which is calculated using the held-out validation set. Furthermore, the image caption C_p for product p generated by $\mathcal{M}_s$ in the captions pool can be expressed as:

$$C_p = \mathcal{M}_s(I_p, A_p, \mathcal{L}_p) \quad (2)$$

To ensure relevance, we discard any caption C_p that does not contain any candidate attribute value from $\mathcal{L}_p$. This filtering step improves the quality of augmentation by eliminating noise and preserves only highly relevant information. The matched attribute-value pairs are extracted as:

$$\hat{L_p} = \{(A_p, V_p) | (A_p, V_p) \in C_p \cap (A_p, V_p) \in \mathcal{L}_p\} \quad (3)$$

Finally, we introduce a large language model (LLM) to summarize all matched captions into a unified context paragraph via $Summarizer(C_p, \hat{L}_p)$. This enriched context is then used to augment the AVE input. In our implementation, we adopt QwenLM as the summarizer due to its strong empirical performance observed in our experiments. The effectiveness of each component in the MICE framework is further analyzed in the ablation study in Section 4.2.2.

3.3 Scalability and Flexibility

The size of the candidate model pool directly affects the computational complexity of the proposed MICE framework. Incorporating a large number of LMMs significantly increases the cost of generating image captions, as GPU and memory consumption scales approximately linearly with the number of models, leading to longer runtimes and higher resource demands. However, the inclusion of the model selection and attribute-value matching components plays a critical role in reducing algorithmic complexity and runtime latency by filtering out underperforming or irrelevant models and captions early in the pipeline.

While it is generally observed that a larger candidate pool yields a greater performance boost, our experiments reveal that even a small subset of strong models (e.g., 1–3) can substantially improve AVE accuracy. In extreme cases, we find that powerful models such as Qwen-VL-Chat benefit significantly from a self-captioning approach, without relying on additional image captioning experts. Detailed empirical results supporting these observations are presented in Section 4.2.2.

Overall, the proposed framework offers flexibility for balancing performance and efficiency, making it adaptable to a wide range of industrial scenarios. For instance, in latency-sensitive online environments, a minimal model pool might work reasonably well and meet real-time SLA requirements, whereas in latency-tolerant offline settings, the full model pool might be leveraged for maximum performance. This adaptability makes the approach well-suited for diverse industrial applications, accommodating varying constraints in terms of resources, latency, and scalability. Note that absolute latency numbers were not compared in this paper, since they depend on multiple factors including device specifications, infrastructure conditions, and network characteristics.

4 Experiments

In this section, we present a comprehensive evaluation of the proposed augmentation approach on the publicly available *ImplicitAVE* dataset (Zou et al., 2024a), a refined multimodal e-Commerce product attributes dataset with five different product categories sourced from MAVE (Yang et al., 2022). We assess the effectiveness of our method in both zero-shot and fine-tuned settings. To further validate its robustness and real-world applicability, we conduct additional experiments on a proprietary women's tops dataset collected from a leading e-commerce platform. These experiments demonstrate the model's performance in a practical deployment scenario. Details of both datasets used in the experiments can be found in Table 1.

Table 1: Dataset Statistics.

Dataset	Category	#Train	#Val	#Test
ImplicitAVE	Clothing	15132	3736	226
	Jewelry	10473	2588	220
	Footwear	17091	4351	317
	Home	9292	2324	457
	Food	2893	724	390
Propriety Dataset	Women Tops	19462	2162	9920

4.1 Experimental Setup

We adopt the same evaluation metric (micro-F1) as used in *ImplicitAVE*. We selected the following SOTA LMM families as benchmarks in the zero-shot setting: BLIP-2 (Li et al., 2023) (Blip2-opt-2.7B, Blip2-flan-t5-xl, Blip2-flan-t5-xxl), InstructBLIP (Dai et al., 2024) (InstructBLIP-vicuna, InstructBLIP-flan-t5), LLaVA (Liu et al., 2024c,a,b) (llava-llama-2, llava-vicuna, llava-v1.6-mistral), InternVL (Chen et al., 2024) (InternVL2-2B, InternVL2-4B, InternVL2-8B), Qwen (Bai et al., 2023; Yang et al., 2024) (Qwen-VL-7B, Qwen-VL-Chat, Qwen2-VL-7B-Instruct). Empirically, across all experiments, we select the three best-performing models on the validation set, InternVL2-4B, Qwen-VL-Chat, InstructBLIP, to construct the LMM candidate pool. Additionally, we compare the fine-tuned results of LAVIN (Luo et al., 2023) and DEFLATE (Zhang et al., 2023b), as reported by (Zou et al., 2024a), as well as GPT-4V (Ouyang et al., 2022), against our finetuned Qwen-VL-Chat and its augmented version.

For fine-tuning Qwen-VL-Chat as the backbone of the MICE framework, we implement the model using PyTorch and optimize it with the Adam optimizer. We adopt the LoRA (Low-Rank Adaptation) technique for efficient parameter tuning. The learning rate is set to 3e-4, with a weight decay of 0.1. Training is conducted with a batch size of 2 per device for 5 epochs. All experiments are conducted on Nvidia A100 GPUs. The prompt template we use follows the same as the prompt used in *Implici-*

Table 2: Experimental results, micro-F1 (%), of an array of selected LMMs using only image (I), image + title (I + T) and mixture of image captioning experts (MICE) across five categories of *ImplicitAVE* for attribute value extraction in a zero-shot setting. Best results of each model for each category are highlighted in bold.

Model	Clothing			Jewelry			Footwear			Home			Food		
	I	I + T	MICE	I	I + T	MICE	I	I + T	MICE	I	I + T	MCIE	I	I + T	MICE
Blip2-opt-2.7b	24.78	21.24	**35.84**	30.45	38.64	**54.09**	19.24	20.19	**35.65**	42.45	42.45	**52.52**	33.33	24.87	**58.97**
Blip2-flan-t5-xl	39.38	30.09	**53.54**	69.09	70.91	**84.55**	44.79	44.79	**59.94**	66.74	68.49	**70.68**	59.49	57.95	**72.05**
Blip2-flan-t5-xxl	46.46	52.65	**67.26**	84.09	**81.82**	**81.82**	56.78	55.84	**64.35**	**72.43**	70.90	**72.43**	73.33	72.31	**77.44**
InstructBLIP-vicuna	**59.29**	42.04	58.41	75.45	75.45	**83.18**	51.10	50.16	**63.09**	60.39	56.67	**67.83**	54.62	55.38	**79.23**
InstructBLIP-flan-t5	48.67	60.18	**67.26**	81.36	**82.27**	79.09	55.84	61.51	**63.72**	72.87	**74.18**	72.87	73.08	75.38	**78.46**
llava-llama-2	20.80	19.47	**53.98**	60.00	63.18	**87.27**	34.07	41.64	**60.25**	59.74	67.18	**68.27**	56.15	56.15	**78.72**
llava-vicuna	20.35	23.01	**51.33**	65.45	60.91	**83.64**	36.91	38.80	**59.94**	61.05	58.42	**67.83**	58.21	43.33	**76.92**
llava-v1.6-mistral	39.82	39.82	**66.37**	74.55	76.36	**89.55**	35.96	44.48	**63.09**	68.27	**72.65**	71.99	73.85	76.67	**84.36**
InternVL2-2B	34.07	33.63	**47.79**	**75.91**	73.64	75.45	32.81	31.86	**47.17**	59.74	61.27	**64.63**	69.49	68.97	**73.85**
InternVL2-4B	45.13	46.90	**66.37**	75.00	76.36	**80.91**	41.64	45.74	**66.04**	63.89	68.05	**75.11**	78.46	78.72	**85.13**
InternVL2-8B	57.52	60.18	**70.80**	76.36	75.91	**80.00**	51.74	57.10	**71.07**	70.02	**72.87**	71.62	80.51	80.00	**85.64**
Qwen-VL-7B	64.16	54.87	**66.37**	85.00	83.64	**88.27**	59.62	56.78	**64.98**	**74.18**	71.99	73.90	75.90	71.54	**85.64**
Qwen-VL-Chat	76.11	69.47	**76.99**	87.27	86.36	**90.00**	68.45	66.25	**72.24**	78.99	**79.65**	78.56	85.13	84.10	**87.18**
Qwen2-VL-7B-Instruct	15.93	26.11	**65.93**	17.27	45.45	**85.00**	14.83	45.74	**64.98**	22.76	56.24	**65.21**	12.05	37.18	**80.77**
Average	42.32	41.40	**60.59**	68.38	70.78	**81.63**	43.13	47.21	**61.19**	62.39	65.79	**69.53**	63.11	63.04	**78.88**

tAVE (Zou et al., 2024a): *"Question: What is the attribute of this product? {mixture of captions}. You must only answer the question with exactly one of the following options {attribute values set}".*

Table 3: Experimental results, micro-F1 (%), of fine-tuned LMMs, GPT-4V, and MICE on *ImplicitAVE* for attribute value extraction. Qwen here is Qwen-VL-Chat.

Model	Clothing	Jewelry	Footwear	Home	Food
DEFLATE*	54.42	67.73	71.61	52.56	61.71
LAVIN*	65.93	78.64	75.39	60.77	64.33
GPT-4V*	77.43	90.45	81.39	**89.93**	90.77
Qwen (finetuned)	82.30	88.64	79.81	83.59	87.69
Qwen (MICE)	**85.40**	**91.82**	**83.60**	87.31	**91.54**

4.2 Results and Discussions

4.2.1 Main Results

Table 2 presents a micro-F1 score comparison for each selected large multimodal model (LMM) under a zero-shot setting, evaluating three input configurations: image only, image + title, and MICE augmented across the five categories of *ImplicitAVE*. The results indicate that the effectiveness of incorporating product/item titles varies significantly across models and categories.There is no single model with a consistent performance advantage. Notably, image-only inputs outperform image + title in the Clothing and Food categories, while image + title provides a slight advantage in Jewelry, Footwear, and Home. The proposed image caption augmentation approach, which only requires a single input modality, further enhances performance by leveraging multiple generated captions, yielding average absolute gains of 14.4 and 12.6 points over image-only and image + title, respectively. This performance boost is consistently observed across most base models, with only a few outliers, demonstrating the robustness and generalizability of the augmentation strategy for zero-shot attribute value extraction.

As Qwen-VL-Chat demonstrated the strongest zero-shot performance, we fine-tuned it into two variants using image-only and image + title data, respectively, and compared it against fine-tuned DEFLATE, LAVIN, and GPT-4V using the micro-F1 metric, as shown in Table 3. Since the training details or checkpoints of DEFLATE and LAVIN are not publicly available and GPT-4V is a closed-source commercial model, we use the results reported by (Zou et al., 2024a) (marked with ∗) in this experiment. Given that (Zou et al., 2024a) only reports results under the image + title configuration, we present the best micro-F1 score between our image-only and image + title fine-tuned Qwen-VL-Chat models. To further enhance performance, we applied the proposed augmentation (MICE) approach, which led to substantial improvements over the fine-tuned Qwen-VL-Chat baseline. The results in Table 3 indicate that fine-tuned Qwen-VL-Chat significantly outperforms DEFLATE and LAVIN, achieving competitive micro-F1 scores comparable to GPT-4V. More importantly, with MICE augmentation, Qwen-VL-Chat surpasses GPT-4V across nearly all categories, except for Home, demonstrating the effectiveness of our approach in augmentating multimodal AVE.

We further evaluate the effectiveness of the proposed augmentation method on a proprietary women's tops dataset from a major e-commerce

Table 4: Experimental results, micro-F1(%) on a proprietary women tops dataset over four target attributes.

	Sleeve	Neckline	Pattern	Color
Image	88.80	52.33	71.47	80.27
Image + Title	84.87	65.80	**81.67**	72.13
Image with MICE	92.07	60.13	73.07	81.13
Image with MICE + Title	**94.40**	**67.00**	78.60	**83.87**

Table 5: Ablation study of MICE over ImplicitAVE.

	Clothing	Jewelry	Footwear	Home	Food
Base	69.47	86.36	66.25	**79.65**	84.10
Majority Voting	66.37	89.09	59.62	79.43	81.03
Self-Captioning	74.78	87.27	69.40	78.34	82.56
w/o (select&match)	67.26	88.18	69.40	72.65	86.64
w/o select	71.68	88.18	68.45	73.96	86.92
w/o summarizer	71.24	**90.00**	68.14	78.34	85.90
ALL	**76.99**	**90.00**	**72.24**	78.56	**87.18**

marketplace, which includes four key product attributes: Sleeve Length, Neckline, Pattern, and Color. Given Qwen-VL-Chat's overall performance from previous experiments, we report micro-F1 scores for fine-tuned Qwen-VL-Chat under four configurations: image only, image + title, image with MICE, and image with MICE + title, as shown in Table 4. Consistent with previous findings, the proposed augmentation approach enhances AVE performance, regardless of whether image-only or image + title inputs are used. This experiment further underscores the real-world applicability and effectiveness of our method for e-commerce attribute value extraction.

4.2.2 Ablation Study

In the previous sections, we have demonstrated the effectiveness of MICE on both a public open-source dataset and a proprietary e-commerce dataset. To better understand the impact of each component, we conduct an ablation study to assess the contribution of each key component (selection, matching, and summarization) to the end-to-end performance. Additionally, we establish baselines using three naive methods for comparison.

As shown in Table 5, each row labeled as ('w/o') reports the micro-F1 score for each product category when a specific component is disabled. By comparing these ablated configurations against the final row (ALL, the complete approach), we observe that selection, matching, and summarization all contribute significantly to overall performance, as their removal results in varying degrees of degradation. Notably, the most substantial performance drop occurs when both selection and matching are disabled, as evidenced by the w/o (select&match) row, highlighting the importance of attribute-aware selection and filtering in our approach.

The three naive baselines considered in this study are: (1) Base, which refers to the fine-tuned Qwen-VL-Chat model without augmentation; (2) Majority Voting, where attribute values are directly extracted from each selected LMM (without generating captions) and aggregated via a majority voting mechanism to determine the final prediction for each attribute; and (3) Self-Captioning, where the fine-tuned Qwen-VL-Chat generates its own image captions for self-augmentation without leveraging external captioning models. As shown in Table 5, when comparing these baselines against the final row (ALL, representing the complete augmentation approach), we observe that the proposed method significantly improves attribute value extraction performance across all categories. These results demonstrate that the proposed augmentation approach effectively enhances attribute value extraction by incorporating multi-source image captioning and attribute-aware selection mechanisms.

4.2.3 Case Study

Figure 1 presents examples of attribute value extraction (AVE) across five product categories in the *ImplicitAVE* dataset under three input configurations: image only, image + title, and image + MICE augmentation, which achieved the best zero-shot performance as seen in Table 2. As shown, the image-only input can lead to incorrect predictions due to subtle or visually ambiguous features. Incorporating the product title does not always help and can introduce misleading information, further degrading model performance. For instance, in the snow boot example, the word "waterproof" in the title causes the model to incorrectly predict "Rain Boot" instead of "Snow Boot". In contrast, MICE-generated captions contain critical context, such as "snow boots", that resolve visual ambiguity and guide the model to the correct prediction. Similarly, in the henley neckline example, the product title lacks key discriminative information, whereas MICE includes the word "buttons", which clearly differentiates a henley from a crew neck. These examples illustrate how MICE enhances the model's understanding by supplementing missing or ambiguous signals from image and text inputs.

To gain deeper insights into the failure modes and attribute-specific weaknesses in MICE, we perform a detailed error analysis as presented

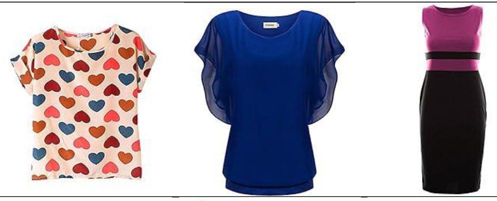
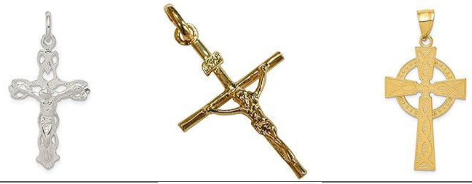
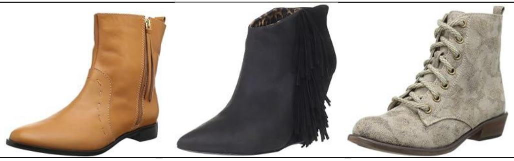
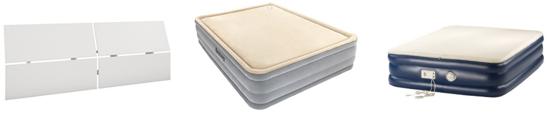
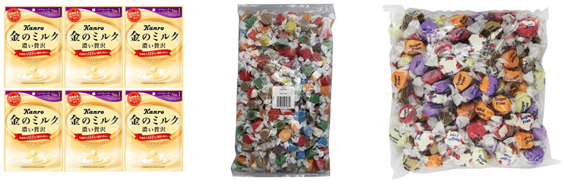
Attribute	Micro-F1	Incorrect Predictions	Example Images
Neckline	69.09%	Label: crew neck Pred: cowl neck	
Shape	84.00%	Label: crucifix Pred: cross	
Shaft Height	50.00%	Label: bootie Pred: ankle boot	
Size	36.67%	Label: queen Pred: full	
Candy Variety	68.29%	Label: taffy Pred: hard candy	

Table 6: Examples of error cases from five categories, highlighting the attribute with the lowest accuracy in each category.

in Table 6. Our observations reveal substantial performance variations among different attributes in some categories using MICE. Specifically, attributes exhibiting lower accuracy typically fall into two categories: 1) Captioning models struggle to capture fine-grained visual details, particularly when certain attributes require contextual references that are absent in the images. This limitation significantly affects the accuracy of caption generation. For instance, accurately identifying specific attributes such as mattress sizes (e.g., full or queen) from a single image without additional context is challenging, leading to inaccuracies in generated captions. 2) Some attributes depend on specialized terminology, which MICE often does not possess such requisite domain knowledge. For example, domain-specific terms such as "taffy", "booties", or "crucifix" have precise meanings within their respective product categories. Without explicit domain expertise, the model struggles to accurately interpret these terms, resulting in erroneous caption generation.

5 Conclusion

In product AVE, scenarios often arise where textual inputs such as product descriptions and titles are either unavailable or unreliable due to ambiguity, incompleteness, or inconsistency. To address this challenge, we propose a novel augmentation framework, Mixture of Image Captioning Experts (MICE), which generates fine-grained, concise, and accurate captions from input images. By leveraging a curated pool of image captioning models, MICE enhances AVE performance, particularly in settings where only visual data is available. Extensive experiments on both the public *ImplicitAVE* dataset and a proprietary women's tops dataset demonstrate that MICE significantly improves the performance of SOTA LMMs in both zero-shot and fine-tuned settings. The modular design of MICE also offers scalability and deployment flexibility, making it suitable for a wide range of industrial use cases with diverse resource and latency constraints.

References

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

Ansel Blume, Nasser Zalmout, Heng Ji, and Xian Li. 2023a. 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.

Ansel Blume, Nasser Zalmout, Heng Ji, and Xian Li. 2023b. Generative models for product attribute extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 575–585.

Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. *CoRR*, abs/1903.12457.

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*, pages 152–159, Singapore. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2024. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Aloïs De la Comble, Anuvabh Dutt, Pablo Montalvo, and Aghiles Salah. 2022. Multi-modal attribute extraction for e-commerce. *arXiv preprint arXiv:2203.03441*.

Zhongfen Deng, Wei-Te Chen, Lei Chen, and S Yu Philip. 2022a. 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)*, pages 1816–1821. IEEE.

Zhongfen Deng, Wei-Te Chen, Lei Chen, and Philip S. Yu. 2022b. 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)*, pages 1816–1821.

Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024. Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2910–2914.

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*, CIKM '23, page 3902–3907, New York, NY, USA. Association for Computing Machinery.

Jiaying Gong, Ming Cheng, Hongda Shen, Pierre-Yves Vandenbussche, Janet Jenq, and Hoda Eldardiry. 2025a. Visual zero-shot E-commerce product attribute value extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 460–469, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiaying Gong, Ming Cheng, Hongda Shen, Pierre-Yves Vandenbussche, Janet Jenq, and Hoda Eldardiry. 2025b. Visual zero-shot e-commerce product attribute value extraction. *arXiv preprint arXiv:2502.15979*.

Jiaying Gong and Hoda Eldardiry. 2024. Multi-label zero-shot product attribute-value extraction. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2259–2270, New York, NY, USA. Association for Computing Machinery.

Jiazhen Hu, Jiaying Gong, Hongda Shen, and Hoda Eldardiry. 2025a. Hypergraph-based zero-shot multimodal product attribute value extraction. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4853–4862. Association for Computing Machinery.

Jiazhen Hu, Jiaying Gong, Hongda Shen, and Hoda Eldardiry. 2025b. Hypergraph-based zero-shot multimodal product attribute value extraction. In *THE WEB CONFERENCE 2025*.

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.

Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johanes Effendi, Justin Chiu, et al. 2024. Rakutenai-7b: Extending large language models for japanese. *arXiv e-prints*, pages arXiv–2403.

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 *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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*, KDD '21, page 3262–3270, New York, NY, USA. Association for Computing Machinery.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

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

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

Shilei Liu, Lin Li, Jun Song, Yonghua Yang, and Xiaoyi Zeng. 2023b. 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*, WSDM '23, page 1039–1047, New York, NY, USA. Association for Computing Machinery.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in neural information processing systems (NeurIPS)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

RetailTouchpoints. 2016. Inconsistent product info spurs returns, erodes customers' trust.

Kalyani Roy, Pawan Goyal, and Manish Pandey. 2024. Exploring generative frameworks for product attribute value extraction. *Expert Systems with Applications*, 243:122850.

Kassem Sabeh, Mouna Kacimi, Johann Gamper, Robert Litschko, and Barbara Plank. 2024. Exploring large language models for product attribute value identification. *arXiv preprint arXiv:2409.12695*.

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*.

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.

Kai Wang, Jianzhi Shao, Tao Zhang, Qijin Chen, and Chengfu Huo. 2023. Mpkgac: Multimodal product attribute completion in e-commerce. In *Companion Proceedings of the ACM Web Conference 2023*, pages 336–340.

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.

Shuhui Wu, Zengming Tang, Zongyi Guo, Weiwei Zhang, Baoliang Cui, Haihong Tang, and Weiming Lu. 2023. Pumgpt: A large vision-language model for product understanding. *arXiv preprint arXiv:2308.09568*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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

Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. 2022. Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1256–1265, New York, NY, USA. Association for Computing Machinery.

Bryan Zhang, Taichi Nakatani, Daniel Vidal Hussey, Stephan Walter, and Liling Tan. 2024a. Don't just translate, summarize too: Cross-lingual product title generation in e-commerce.

Bryan Zhang, Taichi Nakatani, and Stephan Walter. 2024b. Enhancing e-commerce product title translation with retrieval-augmented generation and large language models. *Preprint*, arXiv:2409.12880.

Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023a. 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.

Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023b. 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.

Henry Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip Yu, and Cornelia Caragea. 2024a. ImplicitAVE: An open-source dataset and multimodal LLMs benchmark for implicit attribute value extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 338–354, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Henry Zou, Gavin Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024b. EIVEN: Efficient implicit attribute value extraction using multimodal LLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 453–463, Mexico City, Mexico. Association for Computational Linguistics.