Bridging Modality Gaps in e-Commerce Products via Vision-Language Alignment

Yipeng Zhang*, Hongjun Yu*, Aritra Mandal*, Canran Xu*, Qunzhi Zhou*, and Zhe Wu*
eBay Inc.
{yipezhang, hongjyu, arimandal, canxu, qunzhou, zwu1}@ebay.com

***Abstract*—Item information, such as titles and attributes, is essential for effective user engagement in e-commerce. However, manual or semi-manual entry of structured item specifics often results in inconsistent data quality, errors, and a time-intensive process, particularly for Customer-to-Customer sellers. Generating these descriptions directly from item images offers a promising solution. Existing retrieval-based solutions partially address these issues, but often fall short in capturing fine-grained visual details and handling niche or specialized product categories. To address these challenges, we propose Optimized Preference-Based AI for Listings (OPAL), a novel framework to directly generate schema-compliant, high-quality item descriptions from images using a fine-tuned multimodal-large language model (MLLM). OPAL addresses key challenges in multimodal learning in e-commerce, such as modality gaps and the need for fine-grained contextual understanding. Specifically, OPAL integrates novel data refinement methods—MLLM-Assisted Conformity Enhancement and LLM-Assisted Contextual Understanding—to align textual and visual information while addressing the granularity of contextual understanding. We leverage visual instruction tuning and direct preference optimization to fine-tune MLLM, mitigating hallucination risks and improving model performance across different backbones. Through extensive experiments on real-world e-commerce datasets, we demonstrate that OPAL consistently outperforms baseline solutions in generation quality and completion rates. These results highlight OPAL's effectiveness in bridging the modality gap and elevating the standard of automated listing optimization in e-commerce.**

***Index Terms*—Multimedia Systems, E-commerce, Large Language Models**

I. Introduction

In e-commerce, structured item information are essential for effective search, navigation, and user experience enhancements. These descriptions, comprising an informative **title** and a series of **aspect name-value pairs** (e.g., "Department: Men," "Brand: Nike") are typically provided by sellers during the listing process. However, manual or semi-manual entry often results in inconsistent data quality, with errors, omissions, and irrelevant information. This process is also time-intensive, especially for Customer-to-Customer (C2C) sellers, dissuading them from listing more items. However, the images of items are rich in information and intent, using the images to infer obvious information aids the seller in filling relevant information faster and with more reliability. Using images to populate item information makes the listing creation more of a review process than a data entry process.

* Equal contribution.

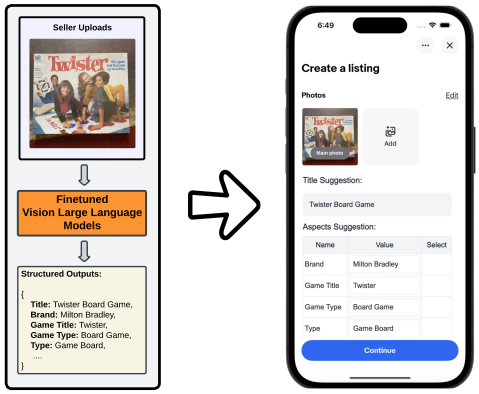

Fig. 1: Overview of the proposed use case. When a user uploads an image, OPAL generates a structured output including title and aspect name-value pairs. This structured output is prefilled for the user, streamlining the listing process and ensuring schema-compliant, high-quality descriptions.

Generating structured descriptions from item images offers a promising solution, but faces significant challenges, especially in C2C scenarios with informal photography, cluttered backgrounds, and varied shooting styles. Although advanced vision models such as YOLO [1] and SAM [2] excel at object detection, they fail to capture the seller's intent or complex attribute relationships. Retrieval-based methods, such as CLIP [3], partially address this by matching images to predefined product candidates, and recent multimodal large language models (MLLM) approach, IPL [4] extends this by generating text based on image retrieval results. However, these inventory-dependent methods often fall short in accurately capturing fine-grained visual details and struggle with less common or specialized product categories due to sparse domain-specific information.

Rather than relying on similarity-based [5] retrieval systems, we propose a novel approach that fine-tunes an MLLM to directly generate structured item descriptions, as illustrated in Figure 1. Our method produces titles and aspects grounded in item images while addressing three critical criteria: (1) Alignment with Seller Intent (capturing the essence of what the seller wants to list), (2) Alignment with Image Information (reflecting visually inferable details such as color or brand), and (3) Conformity to E-commerce Standards (adhering to

predefined schemas for aspects like "Model" or "Publisher"). The proposed method consumes images and extracts visually deductible information, infers relationships, and aligns with knowledge and ontology structure.

Training of MLLM requires high-quality image-text pairs that accurately map product details to their visual context. Although general-purpose datasets like MS COCO [6], Conceptual Captions [7], and Flickr30k [8] are well curated and annotated, e-commerce datasets, especially the product inventory, are inherently noisy and exhibit a significant modality gap. Textual descriptions often include irrelevant information, such as promotional statements, while visual content may inadequately represent key attributes (e.g., storage capacity or dimension). This misalignment increases the risk of hallucination, leading to unreliable outputs. Moreover, the level of granularity required for item recognition in e-commerce far exceeds that of public datasets. Without explicit reinforcement of contextual understanding, the generalizability of the trained MLLM remains limited, compromising its performance in real-world scenarios.

To address these challenges, we developed **O**ptimized **P**reference-Based **AI** for **L**istings (OPAL) equipped with a comprehensive training pipeline that integrates visual and textual signals to produce high quality listings compliant with the schema (Figure 1). Our approach mitigates the modality gap in the e-commerce data while optimizing model performance through visual instruction tuning. The model is further trained with direct preference optimization to enhance contextual understanding. Our contributions can be summarized as follows.

- We analyze the key challenges of multimodal learning in e-commerce, focusing on both the modality gap and the granularity of contextual understanding.
- We propose an effective training strategy (OPAL) to address the modality gap and enhance contextual understanding through refined data and model training. Extensive experiments demonstrate that this strategy consistently improves generation quality across different MLLM backbones.
- We showcase OPAL's superiority over in-house solutions, achieving significant improvements in generation quality and completion rates.

The proposed modality-aligned vision-language model is designed to power an API that extracts and infers structured item information directly from images. These structured outputs can then be applied across various e-commerce applications, such as product listing generation and search relevance enhancement.

II. Related Works

Attribute Value Extraction. Attribute value extraction is crucial in e-commerce for structuring product information from unstructured text. Traditional methods relied on rule-based systems and sequence labeling models like CRFs [9], which demanded extensive manual effort and struggled with scalability. The introduction of transformer-based models, such as BERT [10], significantly improved extraction accuracy by leveraging contextual embeddings. Frameworks like OpenTag [11] introduced pointer networks to handle noisy and fragmented product text. More recently, LLM-based approaches like ExtractGPT [12] have enabled zero-shot and few-shot extraction, demonstrating strong generalization to unseen attribute values. Multilingual models such as GAVEL [13] further extend this capability to diverse markets and languages.

Large Language Models (LLMs) Large Language Models, including GPT-4 [14], PaLM [15], and LLaMA [16], have transformed natural language understanding and generation tasks. In e-commerce, LLMs have been adapted for product categorization, attribute-value generation, personalized recommendations, and listing optimization [17], [18]. Domain-specific LLMs like e-LLaMA [19] improve performance by incorporating proprietary knowledge while preserving data security. Strategies involving schema alignment and user intent modeling further increase the utility of LLMs in structured prediction and personalization tasks [20]–[22].

Multimodal Large Language Models (MLLMs). MLLMs integrate visual and textual modalities to enable comprehensive multimodal reasoning. Early models such as CLIP [3] and ALIGN [23] introduced contrastive learning frameworks to align image and text embeddings for robust zero-shot transfer. More advanced MLLMs like LLaVA [24] QwenVL [25] and InternVL [26] combine visual encoders with LLMs to enable tasks such as visual question answering, image-conditioned generation, and multimodal dialogue. These models are especially relevant in e-commerce scenarios where both product images and textual descriptions are available and complementary.

MLLMs for E-commerce Applications. MLLMs have increasingly been applied to e-commerce for multimodal product understanding, including title rewriting, attribute extraction, and visual-grounded content generation. For instance, models like PUMGPT [27] and VL-GPT [28] leverage visual context to improve contextual understanding. E-commerce-specific MLLMs such as IPL [4] integrate product images, attributes, and category information through image retrieval of the e-commerce database. However, challenges persist due to noisy input data, weak alignment between visual and textual modalities, and inconsistency in attribute schemas [29]. Our work addresses these limitations by improving data alignment and leveraging MLLMs in a structured, schema-aware fashion.

III. Key Challenges in Multimodal Alignment and Contextual Understanding in E-commerce

Modality Gap The modality gap is well-acknowledged in multimodal learning [30]–[32] and refers to the inherent differences in representation, structure, and distribution across modalities like text, image, audio, or video. This gap poses significant challenges for e-commerce datasets, where textual descriptions and image representations often lack alignment, which can result in hallucinations, where models generate plausible but incorrect information due to ambiguous or missing visual representation. Key challenges include: **1) Limi-**

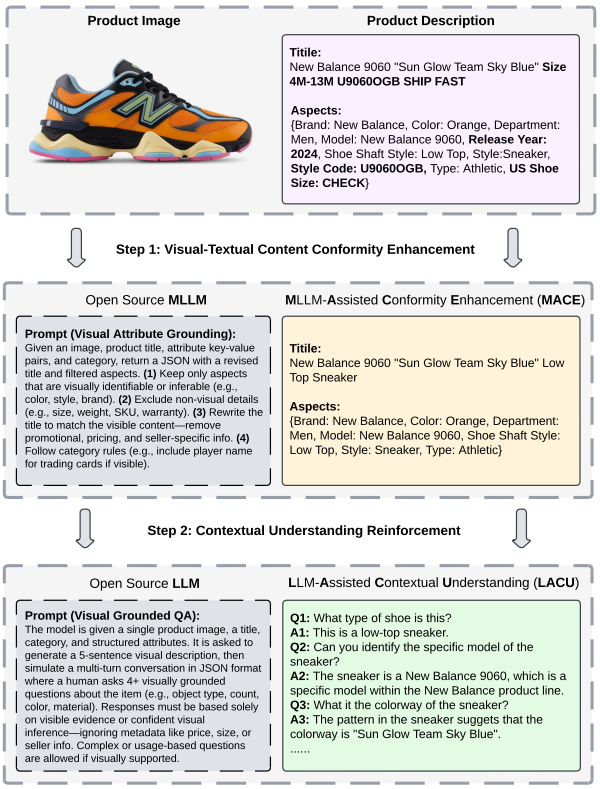

Fig. 2: Overview of the modality gap in e-commerce product data and the proposed MACE–LACU pipeline. The top row illustrates the modality gap commonly observed in e-commerce listings: text descriptions often include information not visually grounded in the image, such as "Size 4M-13M" and seller-specific claims like "SHIP FAST." Training vision-language models on such misaligned data can lead to hallucinations during inference (see Figure 3), as visually unverifiable details could vary across listings. To mitigate this, each image–text pair is processed by MACE to filter out non-visual attributes and rewrite the product title based on image-grounded features. Then, LACU generates multi-turn, visually grounded conversations that reinforce item-specific contextual understanding. The full pipeline consumes an item image, title, category, and structured attributes, and outputs a harmonized, vision-consistent representation for improved multimodal understanding. Both prompts used in MACE and LACU are simplified due to space constraints.

tations of Visual Content: Many product attributes are not visually discernible, such as precise specifications (e.g., "*120 cm × 60 cm × 75 cm*" for furniture dimensions), materials (e.g., "*100% cotton*" for clothing), or unique identifiers (e.g., "*AB1235-567*" for sneakers). Safety certifications and technical details like processor types or battery capacities also can not be represented by images solely unless with explicit textual descriptions in the image. **2) Text-Visual Misalignment:** E-commerce listings often include irrelevant or distracting textual elements, such as promotional content (e.g., "*50% Off!*"), shipping details, or decorative emojis. These details are frequently misaligned with visual content, leading to challenges in training MLLMs. We demonstrate such modality gap using e-commerce example data in Figure 2.

Contextual Understanding Gap in MLLM Pretraining Most MLLMs are pretrained on open-source vision-language datasets such as MS-COCO [6], Conceptual Captions [7], and Open Images [33]. While effective for general-purpose multimodal understanding, these datasets lack the contextual specificity and attribute granularity required for e-commerce applications. Their captions tend to be surface-level—e.g., "a man riding a bike" or "a red handbag on a table"—and rarely capture fine-grained product attributes critical to structured commerce systems. Examples include shoe models (e.g., "Nike Dunk Low," "New Balance 550"), collectible brands (e.g., "Kup Stax," "Funko Pop!," "Bearbrick"), toy character names (e.g., "Grogu (Baby Yoda)," "Optimus Prime," "Elsa from Frozen"), or author names in collectible books (e.g., "K.T. Oslin," "Haruki Murakami"). These attributes are essential for item selling in e-commerce. As a result, MLLMs pretrained on generic datasets often struggle to generalize to commerce-specific tasks that demand precise visual-textual grounding and context-sensitive reasoning.

IV. Optimized Preference-Based AI for Listings

To address the challenges of multimodal learning in e-commerce, we propose the **O**ptimized **P**reference-Based **AI** for **L**istings (OPAL) model. OPAL integrates data enhancement, visual instruction tuning, and preference optimization to bridge modality gaps and improve contextual understanding. Its data pipeline comprises two components: **MLLM-Assisted Conformity Enhancement (MACE)** and **LLM-Assisted Contextual Understanding (LACU)**. MACE aligns visual and textual data by refining noisy e-commerce inputs, while LACU enriches context through domain-specific, multi-turn conversations. As illustrated in Figure 2, these methods produce well-aligned, context-rich training data. Visual instruction tuning is further improved using Direct Preference Optimization (DPO), enabling the model to generate accurate, schema-compliant titles and aspects for e-commerce listings.

A. MLLM-Assisted Conformity Enhancement

To address the modality gap in e-commerce data, we introduce MACE, a preprocessing pipeline that leverages a very large MLLM, InternVL2.5-78B, to refine the alignment between visual and textual modalities. The key idea is to enforce *visual conformity* by filtering out textual information that cannot be visually confirmed or reasonably inferred from the image. Given a product image and its associated description (including title and structured aspects), the MLLM is prompted to perform two operations: (2) it rewrites the title by removing tokens that are not grounded in visual evidence; (2) it drops aspect key–value pairs that cannot be confirmed or inferred from the image.

A simplified example of this prompting strategy is illustrated in Figure 2, Step 1. MACE process removes irrelevant or

TABLE I: Demonstration of the procedure for generating preference pairs used in DPO training. An MLLM (InternVL2.5-78B) evaluates the generated title and aspects against the visually and textually aligned item information using an evaluation prompt. If the generation is judged as *Incorrect* or *Mostly Incorrect*, the item information (chosen) and model output (rejected) form a preference pair for DPO training. In this example, the model incorrectly treats a golf head cover as an animal toy (highlighted).

Image	Example (Simplified) Judge Prompt	MACE Aligned (Chosen)	Model Output (Rejected)	Judge
	**Task**: Evaluate how accurately the predicted title and aspects match the {GOLDENTITLE} and {GOLDENASPECTS}. **Visual Check**: Assess visible properties in the image (color, shape, text, logos). **Aspect Check**: Match predicted aspects with visible properties and labeled aspects. **Title Check**: Check if the predicted title matches the item. **Judgment**: Label *each aspect* as "Correctly Identified" or "Not Correctly Identified" and the title as "Correctly Matched" or "Not Correctly Matched." **Final Evaluation Criteria**: Correctly Generated (95-100%) Mostly Correctly Generated (80-94%) Partially Correctly Generated (50-79%) Mostly Incorrectly Generated (30-49%) Incorrectly Generated (¡30%)	**Title:** Sloth Golf Wood Headcover Fit Driver Fairway Woods Club Plush Head Covers **Aspects:** {Brand : Unbranded, Sport/Activity : Golf, Type : Wood Head Covers, Vintage : No}	**Title:** Sloth Plush Doll Stuffed Animal Toy **Aspects:** {Brand : Unbranded, Character : Sloth, Material : Plush, Theme : Animation, Type : Action Figure }	Incorrectly Generated

unverifiable details, such as shoe size, style code, or seller-specific information like "SHIP FAST". This results in a conformity-enhanced description that is tightly coupled with the visual content. By training downstream models on these visually grounded inputs, MACE enforces a stricter visual-to-text mapping and significantly reduces the modality gap.

B. LLM-Assisted Contextual Understanding

Although MACE effectively aligns text and image modalities, it does not fully capture the granular contextual nuances essential for e-commerce, many of which fall outside the scope of standard multimodal LLM pretraining. To address this limitation, LACU enhances MACE by generating conversational datasets using the textual LLM, LLaMA 3.1-Instruct [34]. Inspired by common practices in visual instruction tuning [24], we utilize this language model to simulate multi-turn dialogues that emulate interactions between a seller and a prospective buyer. These conversations are intentionally constructed to cover a broad (as many as possible) range of product aspects, thereby enriching the model's contextual understanding. A simplified example of the prompt used to query the textual LLM is shown in Figure 2 - Step 2. The synthetic dialogues are designed to span diverse domain-specific queries and detailed product descriptions, equipping the model to achieve a finer-grained understanding of item semantics from the image.

C. Visual Instruction Tuning with DPO

We fine-tune our MLLM backbones using a visual instruction tuning on datasets derived from MACE and LACU. For each image-instruction pair, the model is optimized to minimize the loss on the generated outputs, thereby learning to produce accurate and contextually grounded product titles and attributes based on visual input. Specifically, for each (image-instruction, response) pair (x, y), the model is trained to minimize the negative log-likelihood of the output sequence using standard supervised fine-tuning:

$$\mathcal{L}_{\text{Visual Instruction Tuning}} = -\sum_{t=1}^{T} \log P_\theta(y_t \mid y_{<t}, x) \tag{1}$$

We then apply Direct Preference Optimization (DPO) [35] to further enhance the model's contextual understanding and reduce hallucinations. Specifically, after visual instruction tuning, we judge the model's generation on a subset of the training data using a large vision-language model, InternVL2.5-78B. This judge model receives the image, the ground truth product description, and the model-generated description, and classifies the generation into one of five categories: *Correctly Generated*, *Mostly Correctly Generated*, *Partially Correctly Generated*, *Mostly Incorrectly Generated*, or *Incorrectly Generated*. Generations deemed to have incorrect contextual understanding (labeled as *Mostly Incorrectly Generated* or *Incorrectly Generated*) are used to construct preference pairs for DPO training, where the ground truth is treated as the *chosen* output and the model prediction as the *rejected* output. By creating the pair $(x, y_{\text{chosen}}, y_{\text{rejected}})$. The sigmoid-based DPO loss encourages the model to assign a higher likelihood to the preferred output:

$$\mathcal{L}_{\text{pref}} = -\log \sigma\left(\beta\left[\log \pi_\theta(y_{\text{chosen}} \mid x) - \log \pi_\theta(y_{\text{rejected}} \mid x)\right]\right) \tag{2}$$

A KL-regularization term is included to constrain the updated model to stay close to a reference policy:

$$\mathcal{L}_{\text{DPO}} = \mathcal{L}_{\text{pref}} + \lambda \cdot \text{KL}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right), \tag{3}$$

where π_θ is the current policy, π_{ref} is the reference policy, $\sigma(\cdot)$ is the sigmoid function, β is a temperature parameter controlling preference sharpness, and λ weights the optional

KL regularization term. A simplified illustration of this procedure, along with an example prompt, is provided in Table I.

V. Experiments

A. Experiment Setup

1) Training and Evaluation Datasets:: We collect data comprising millions of product entries from a world-leading e-commerce platform. After an extensive cleaning process, we retain one million valid image–description pairs (titles and aspects). For resource efficiency, only the main image is retained for each product.

Task-Specific Instruction Dataset: Using MACE, we align the data with product images and textual description, ensuring conformity to the e-commerce schema. This process results in 890K high-quality image–instruction pairs in JSON format, each containing a refined title and corresponding aspect-value pairs.

Contextual Understanding Dataset: To enhance contextual understanding, we apply LACU to generate approximately 800K image–conversation pairs. Each pair simulates multi-turn buyer-seller interactions, typically involving at least five conversational rounds. These dialogues provide a rich source of domain-specific and contextually diverse training data. Both the instruction and conversational datasets are merged during training.

Preference Optimization Dataset: We use a large MLLM to evaluate each backbone's responses on a subset of the training data (200K). We then select 20K preference pairs for each backbone, each containing an instruction paired with a preferred (chosen) and a dispreferred (rejected) response.

Evaluation Dataset: We collect an independent dataset of 100K diverse samples comprising images and textual information (titles and aspects) from the e-commerce inventory. These samples undergo human evaluation to ensure accurate ground truth, serving as the benchmark for all downstream analysis. Notably, the dataset contains niche items that appear only after the model has been trained, providing a meaningful test of generalization.

2) Implementation Details: We adopt a two-stage training procedure consisting of visual instruction tuning followed by DPO. All experiments are conducted using the `LLaMA-Factory` [36] framework with full-parameter fine-tuning. The vision encoder is frozen throughout. Training is performed on 8×A100 GPUs (80GB) using DeepSpeed ZeRO-3 [37] and `bfloat16` precision. For visual instruction fine-tuning, the maximum token length is set to 4,096. We use a per-device batch size of 2 with gradient accumulation over 16 steps. Models are trained for 1 epoch using AdamW with a learning rate of 1×10^{-5}, cosine learning rate decay, and a 10% warmup ratio. For DPO, the pairwise preference dataset is formatted with the same prompt template. The model is optimized using a sigmoid-based DPO loss. Training is conducted for 1 epoch with a per-device batch size of 1 and gradient accumulation over 8 steps. We use a learning rate of 5×10^{-6} and the same scheduler as in the previous stage.

TABLE II: Performance comparison of different MLLM backbones and training strategies. An in-context learning evaluation for InternVL2.5-78B is also included for reference.

Backbone	Strategy			Rouge-L F1	Aspect F1	Schema Recall
	MACE	LACU	DPO			
InternVL2.5-78B	In-context			0.25	0.30	0.46
LLaVA-NeXT-7B	-	-	-	0.36	0.33	0.49
	✓	-	-	0.43	0.35	0.53
	✓	✓	-	0.48	0.37	0.54
	✓	✓	✓	**0.56**	**0.49**	**0.74**
Qwen2-VL-7B	-	-	-	0.39	0.38	0.53
	✓	-	-	0.48	0.40	0.56
	✓	✓	-	0.50	0.42	0.56
	✓	✓	✓	**0.61**	**0.50**	**0.85**
InternVL2.5-8B	-	-	-	0.42	0.45	0.71
	✓	-	-	0.52	0.48	0.61
	✓	✓	-	0.53	0.50	0.62
	✓	✓	✓	**0.63**	**0.52**	**0.82**

B. Quantitative Evaluation

1) Evaluation Metrics: We assess the quality of each generated *title* + *aspect*–JSON string using one holistic metric and two targeted metrics that focus on aspect-level correctness. These collectively evaluate both the fluency of the structured output and its alignment with e-commerce expectations.

ROUGE–L F1. To evaluate the overall quality of the generated JSON string as a complete entity, we use ROUGE–L F1, which captures both contextual coherence and structural fidelity. Given the predicted token sequence s^{pred} and the reference s^{ref}, we first compute the length l of their longest common subsequence (LCS). Then:

$$P_{\text{LCS}} = \frac{l}{|s^{\text{pred}}|}, \quad R_{\text{LCS}} = \frac{l}{|s^{\text{ref}}|},$$

$$\text{ROUGE–L F1} = \frac{2P_{\text{LCS}}R_{\text{LCS}}}{P_{\text{LCS}} + R_{\text{LCS}}}.$$

Since the entire structured string is compared, this metric rewards both semantic accuracy and adherence to the expected format.

Aspect-Matching F1. To further evaluate the model's performance on aspect prediction, we extract and normalize all aspect name–value pairs from both the prediction and the reference. Let $\mathcal{A}^{\text{pred}}$ and $\mathcal{A}^{\text{ref}}$ denote the predicted and reference aspect sets, respectively. We define:

$$P_{\text{asp}} = \frac{|\mathcal{A}^{\text{pred}} \cap \mathcal{A}^{\text{ref}}|}{|\mathcal{A}^{\text{pred}}|}, \quad R_{\text{asp}} = \frac{|\mathcal{A}^{\text{pred}} \cap \mathcal{A}^{\text{ref}}|}{|\mathcal{A}^{\text{ref}}|},$$

$$\text{Aspect F1} = \frac{2P_{\text{asp}}R_{\text{asp}}}{P_{\text{asp}} + R_{\text{asp}}}.$$

This fine-grained metric reflects how accurately the model identifies and reproduces key attribute pairs, which is central to downstream applications.

Schema-Compliance Recall. Finally, we evaluate whether the predicted aspect keys conform to the platform's predefined schema. This is critical for ensuring that the outputs can be directly consumed by the e-commerce backend. Let $\mathcal{S}$ be the

set of allowed schema keys and K^{pred} the predicted keys. Then:

$$\text{Schema}_{\text{Recall}} = \frac{|\{k \in K^{\text{pred}} \mid k \in \mathcal{S}\}|}{|K^{\text{pred}}|}.$$

This metric reports the proportion of predicted keys that align with the schema, reflecting practical deployability in product listings.

2) Effectiveness of OPAL Generalizes Across MLLM Backbones: We evaluated OPAL using different MLLM backbones, LLaVA-NeXT-7B [38], Qwen2-VL-7B [25], and InternVL2.5-8B [39], on four variants: **(1) Baseline:** Raw product knowledge without conformity alignment, **(2) MACE:** Data refined to align textual and visual information, **(3) MACE + LACU:** Further enriched data capturing detailed product context, and **(4) MACE + LACU + DPO:** Enhanced contextual understanding through Direct Preference Optimization. As shown in Table II, the proposed methods collectively improved generation quality, enhanced contextual understanding, and optimized overall performance across all backbones. Specifically, models trained with the full OPAL pipeline (MACE + LACU + DPO) achieve at least 50% improvement in Rouge-L F1 (0.63 vs. 0.42) and at least 16% improvement in Aspect F1 (0.52 vs. 0.45) compared to models trained on unaligned data. While higher scores consistently reflect improved generation quality, we do not expect metrics like Rouge-L F1 or Aspect F1 to reach 1.0, as many items can have multiple valid titles and aspect combinations. However, in this context, improvements from moderate to higher scores are still meaningful. Even though fine-tuned InternVL2.5-8B, using unaligned data, shows good Schema Recall (aspect name), the low Rouge F1 and Aspect F1 demonstrate significant hallucination in the aspect value. We also compared with the common industrial practice, in-context learning, using a much larger backbone (InternVL2.5-78B), using category-specific demonstration examples. Despite the larger model size, in-context learning showed significantly lower generation quality, underscoring the advantages of OPAL's fine-tuning approach for improving contextual understanding.

C. Qualitative Evaluation

1) Effectiveness of the MACE and LACU Pipeline.: We qualitatively illustrate the effectiveness of the MACE and LACU pipeline by visualizing generation outputs at different stages of the data refinement process in Figure 3. We evaluate an MLLM (InternVL2.5-8B) trained under varying data preparation regimes, paralleling the quantitative trends reported in Table II. When trained on raw, unaligned descriptions (Baseline in Sec.V-B2), the model produces a significant number of hallucinated attributes, limiting its applicability in real-world settings. While applying MACE helps reduce hallucinations by aligning image-text pairs (MACE), the model remains limited in extracting fine-grained contextual cues, such as specific shoe models, which are often underrepresented or missing in pretraining corpora. Upon further incorporating visual-grounded Q&A supervision through LACU (MACE + LACU), the model demonstrates enhanced contextual comprehension and more accurate visual attribute generation. The results underscore the effectiveness of both MACE and LACU.

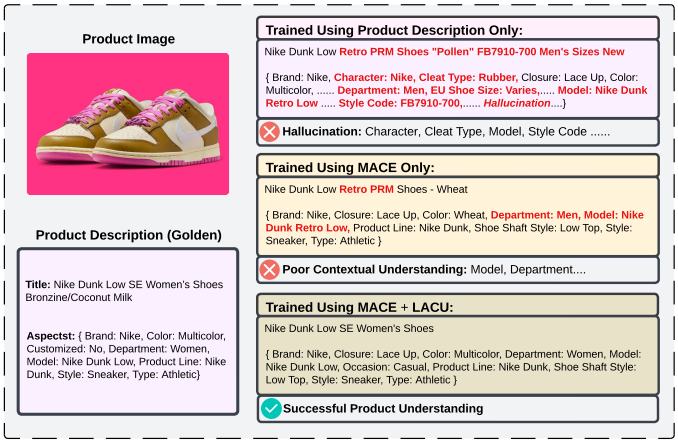

Fig. 3: Qualitative comparison of the MLLM (InternVL2.5-8B) generated from different stages of the training (hallucination is highlighted in red).

2) OPAL Result Visualizations.: We further demonstrate the generations from the OPAL framework (InternVL2.5-8B) in Table III. These cases highlight the model's ability to accurately infer seller intent, extract fine-grained product aspects, and generalize to niche items that fall outside its training distribution. Notably, the first three examples were released after April 2024 and were absent from the training data, which was limited to sources available prior to March 2024. As a result, image similarity-based retrieval methods would be ineffective, given the lack of indexed product knowledge.

In the first example, the key identifying features, namely the Coca-Cola and Paris Olympic icons, appear at the center of an otherwise predominantly white tote bag. Retrieval-based methods, even when the exact same bag is present in the inventory but photographed from slightly different angles or with varied shooting styles, tend to match such images to other generic white tote bags. This often results in incorrect product identification. The third example highlights OPAL's ability to capture fine-grained details, such as the PSA certification number, card grade, and unique identifier. Although these elements occupy only a small portion of the image, they are critical and highly relevant to sellers. OPAL not only identifies these details but also infers contextual information, such as the card manufacturer and its parallel/variety, even when such information is not explicitly visible in the image. These capabilities demonstrate OPAL's strong contextual reasoning and its potential to enhance product understanding in e-commerce. The last two examples demonstrate OPAL's robustness in handling ambiguous images with limited visual cues regarding seller intent. Retrieval-based methods frequently misidentify these items by matching them to visually similar but semantically incorrect products. In contrast, OPAL exhibits strong domain-specific reasoning, generating accurate titles and aspect values that align with the true intent of the seller.

TABLE III: Model outputs from the OPAL framework using InternVL2.5-8B. Each row shows the input image, generated title, and predicted aspects, highlighting the model's ability to interpret niche items and seller intent.

#	Image	Model Output (Title + Aspects)
1		**Title:** Coca-Cola 2024 Paris Summer Olympics Tote Bag **Aspects:** {Brand: Coca-Cola, Color: White, Gender: Unisex Adult, Modified Item: No, Officially Licensed: Yes, Product: Tote Bag, Sport: Soccer, Team: Olympics}
2		**Title:** Pokemon x Le Creuset Pokemon Special Edition Coffee Mug Set **Aspects:** {Brand: Le Creuset, Type: Coffee Mug Set, Theme: Pokemon, Material: Ceramic, Color: Multicolor, Number of Items in Set: Five-Piece, Features: Dishwasher Safe}
3		**Title:** 2024 Donruss Optic Preview Pink Pulsar Rated Rookie RC #303 J.J. McCarthy **Aspects:** {Features: Rookie, Manufacturer: Panini, Parallel/Variety: Pink, Season: 2024, Set: 2024 Donruss Optic, Sport: Football, Type: Sports Trading Card, Player: J.J. McCarthy, Team: Minnesota Vikings, Card Number: 303, Grade: PSA 10 GEM MT, PSA Certification Number: 108300232}
4		**Title:** Solar LED Landscape Pathway Lights, 2 Lights, Dusk to Dawn, Waterproof, Garden **Aspects:** {Brand: Unbranded, Bulb Type: LED, Features: Dusk-to-Dawn, Waterproof, Power Source: Solar, Style: Art Deco, Type: Deck/Step Light}
5		**Title:** Soft Silicone Phone Case Cover with Lanyard **Aspects:** {Brand: Unbranded, Compatible Brand: For Apple, Design/Finish: Luxury, Features: Shockproof, Material: Silicone/Gel/Rubber, Type: Back Case}

TABLE IV: Offline comparison with baseline model

	Final Submission Rate	Conformity	Win Rate
OPAL (InternVL2.5-8B)	+53.40%	+10.50%	60%

D. Comparison with a Baseline System

To further evaluate the performance of OPAL, we compared it with a similarity-based baseline system leveraging multimodal retrieval. This pipeline employs a large-scale multimodal embedding model with a knowledge graph containing over one billion edges, meticulously developed over an 18-month period. The independent evaluation dataset comprises thousands of actual listing sessions, each featuring high-quality human annotations for titles and aspects. We employed three metrics for this comparison: **Final Submission Rate**: Assesses how effectively the model helps sellers complete listings when images are uploaded using an in-house user traffic simulation method. **Conformity**: Evaluated using a large MLLM (InternVL2.5-78B) to measure alignment with ground truth with alignment score from 1 to 5, similar as that in Table I, where 5 represents correct generation and 1 represents incorrect generation. **Win Rate**: Direct comparison with the baseline system, rated by the same large MLLM to determine if generation from the OPAL wins that from the baseline model. As demonstrated in Table IV, OPAL outperforms the baseline system by generating higher-quality titles and aspects, as well as significantly improving the final submission rate. These enhancements showcase a more simplified and efficient listing experience.

VI. Conclusion

In this work, we introduced **O**ptimized **P**reference-Based **AI** for **L**istings (OPAL), a novel vision-language framework designed for e-commerce listing optimization. OPAL generates high-quality item information, including titles and structured aspects, directly from product images. It addresses key challenges in multimodal learning, such as the modality gap and lack of attribute-level context in LLM pretraining, through our proposed pipeline: (1) MACE filters and aligns noisy text with image-grounded data; (2) LACU injects contextual knowledge via visually grounded Q&A generation; and (3) DPO fine-tunes outputs toward better contextual understanding. Experiments show OPAL outperforms strong baselines across multiple backbones and metrics, generalizing well to unseen or niche items. It effectively infers subtle product traits and seller intent, demonstrating strong contextual reasoning from multimodal data.

While alternative methods such as guided decoding using a very large MLLM or agent-based retrieval systems offer complementary value, they typically require extensive external infrastructure and are resource-intensive. Furthermore, they still fall short in bridging the underlying modality gap and contextual reasoning limitations of pretrained models. By contrast, OPAL offers a unified solution that embeds these capabilities directly through end-to-end fine-tuning, and it can still be incorporated into these advanced frameworks.

OPAL is developed as an API that enables high-quality, schema-aligned product information from images, supporting not only listing assistance but also multimodal tasks like image interpretation, product content understanding, and automated labeling. It accelerates training data annotation and has broader applications across e-commerce. Looking ahead, we plan to add multi-image reasoning and expand OPAL into an agentic framework that can query external knowledge, interact with sellers, and adapt to category-specific constraints. These advances will move OPAL toward becoming an end-to-end multimodal assistant, highlighting the transformative potential of fine-tuned vision-language models in automating and scaling listing generation.

REFERENCES

[1] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[3] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[4] K. Chen, Q. Zhang, C. Lian, Y. Ji, X. Liu, S. Han, G. Wu, F. Huang, and J. Chen, "Ipl: Leveraging multimodal large language models for intelligent product listing," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, 2024, pp. 697–711.

[5] A. Mandal, D. Tunkelang, and Z. Wu, "Semantic equivalence of e-commerce queries," 2023. [Online]. Available: https://arxiv.org/abs/2308.03869

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[7] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[8] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[9] B. P. Majumder, N. Potti, S. Tata, J. B. Wendt, Q. Zhao, and M. Najork, "Representation learning for information extraction from form-like documents," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 6495–6504.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] J. Zheng *et al.*, "Opentag: Open attribute value extraction from product titles and descriptions," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[12] A. Brinkmann, R. Shraga, and C. Bizer, "Extractgpt: Exploring the potential of large language models for product attribute value extraction," in *International Conference on Information Integration and Web Intelligence*. Springer, 2025, pp. 38–52.

[13] P. Hongwimol, D. Sheng, L. Zhang, K. Liu, and X. Wang, "Gavel: Generative attribute-value extraction using llms on llm-augmented datasets," in *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, 2025, pp. 81–90.

[14] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[15] A. Chowdhery *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[16] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2307.09288*, 2023.

[17] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, K. Zheng, D. Lian, and E. Chen, "When large language models meet personalization: perspectives of challenges and opportunities," *World Wide Web*, vol. 27, no. 4, p. 42, 2024.

[18] L. Li, Y. Zhang, and L. Chen, "Personalized transformer for explainable recommendation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 4947–4957.

[19] C. Herold, M. Kozielski, T. Bazazo, P. Petrushkov, S. H. Hashemi, P. Cieplicka, D. Basaj, and S. Khadivi, "Domain adaptation of foundation llms for e-commerce," *arXiv preprint arXiv:2501.09706*, 2025.

[20] S. Zhang, B. Peng, X. Zhao, B. Hu, Y. Zhu, Y. Zeng, and X. Hu, "Llasa: Large language and e-commerce shopping assistant," *CoRR*, vol. abs/2408.02006, 2024.

[21] B. Peng, X. Ling, Z. Chen, H. Sun, and X. Ning, "ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data," in *Forty-first International Conference on Machine Learning*. OpenReview.net, 2024.

[22] C. Palen-Michel, R. Wang, Y. Zhang, D. Yu, C. Xu, and Z. Wu, "Investigating LLM applications in e-commerce," *CoRR*, vol. abs/2408.12779, 2024.

[23] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 980–17 989.

[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916.

[25] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[26] K. Chen *et al.*, "Internvl: Advancing vision-language alignment with large-scale pre-training," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[27] W. Xue, Z. Guo, B. Cui, Z. Xing, X. Zeng, X. Wang, S. Wu, and W. Lu, "Pumgpt: A large vision-language model for product understanding," *arXiv preprint arXiv:2308.09568*, 2023.

[28] J. Zhu, X. Ding, Y. Ge, Y. Ge, S. Zhao, H. Zhao, X. Wang, and Y. Shan, "Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation," *arXiv preprint arXiv:2312.09251*, 2023.

[29] Z. Hu, S. Li, M. Du, A. Dhua, and D. Gray, "De-noised vision-language fusion guided by visual cues for e-commerce product search," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE, 2024, pp. 1986–1996.

[30] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6787–6800. [Online]. Available: https://aclanthology.org/2021.emnlp-main.544/

[31] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *NeurIPS*, 2022. [Online]. Available: https://openreview.net/forum?id=S7Evzt9uit3

[32] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 2–25.

[33] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International journal of computer vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[34] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[35] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: your language model is secretly a reward model," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2024.

[36] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma, "Llamafactory: Unified efficient fine-tuning of 100+ language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Bangkok, Thailand: Association for Computational Linguistics, 2024. [Online]. Available: http://arxiv.org/abs/2403.13372

[37] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: Memory optimizations toward training trillion parameter models," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2020, pp. 1–16.

[38] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," 2024.
[39] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.