# MOON2.0: Dynamic Modality-balanced Multimodal Representation Learning for E-commerce Product Understanding

Supplementary Material

## 7. Model Architecture and Internal Tools

Our backbone is a **bi-encoder vision-language model** comprising the following components: The **language encoder** is a 24-layer Transformer with 24 attention heads, a hidden size of 3072, and a vocabulary size of 129,280. The **vision encoder** utilizes a 32-layer OpenCLIP-based ViT featuring 16 heads, a hidden size of 1280, a patch size of 14, and an input image resolution of $378 \times 378$.

For product entity extraction, we initially employ an internal BERT-based Named Entity Recognition (NER) model. This model is fine-tuned on 500,000 annotated e-commerce titles, achieving a 92% F1 score in identifying brands, categories, and materials. To facilitate public reproducibility, we demonstrate that this proprietary tool can be replaced with a pipeline combining spaCy and rule-based filtering, which results in a marginal performance drop of $<2\%$ in R@10 on the *MBE2.0* retrieval tasks.

## 8. Results of Finetuned MLLM

For equitable comparison, Qwen2.5-VL and InternVL3 were fine-tuned on the *MBE2.0* training set. Table 5 confirms the architectural and methodological contributions of *MOON2.0*.

Table 5. Retrieval results (R@1) on *MBE2.0* before and after SFT.

| Methods \ Metrics | $q^{t} \rightarrow c^{mm}$ | | $q^{i} \rightarrow c^{mm}$ | | $q^{mm} \rightarrow c^{mm}$ | |
|---|---|---|---|---|---|---|
| | 0-shot | sft | 0-shot | sft | 0-shot | sft |
| InternVL3-2B | 0.11 | 23.14 | 2.58 | 33.70 | 4.76 | 35.03 |
| Qwen2.5-VL-3B | 1.81 | 24.28 | 13.04 | 31.23 | 18.50 | 34.20 |
| ***MOON2.0* (Ours)** | - | **27.34** | - | **41.07** | - | **43.34** |

## 9. Retrieval Results

To further evaluate the retrieval efficacy and generalization capacity of our *MOON2.0* across diverse e-commerce scenarios, we conducted large-scale retrieval experiments on a randomly sampled retrieval set comprising five million samples drawn from the platform's comprehensive search database spanning the period from July 1, 2025, to the present.

This evaluation encompasses three retrieval scenarios: (i) text-to-multimodal retrieval ($q^{t} \rightarrow c^{mm}$), (ii) image-to-multimodal retrieval ($q^{i} \rightarrow c^{mm}$), and (iii) multimodal-to-multimodal retrieval ($q^{mm} \rightarrow c^{mm}$), which are expected to provide a relatively comprehensive evaluation of the model's retrieval capabilities. All queries employed in these experiments are sourced from authentic user data.

As presented in Fig. 8, the $q^{t} \rightarrow c^{mm}$ retrieval results exhibit strong alignment between query semantics and retrieved items. For instance, in the top row of the figure, the product characterized by the "Cream-colored Bear" and "Bag Charm" features occupies the Top-1 position (Rank 1). Its accompanying description prominently highlights attributes such as "Cream-colored Bear", "Brown Overalls" and "Plush", underscoring *MOON2.0* 's precise interpretation of query intent and its high fidelity in single-item retrieval. Notably, even as the retrieval scope expands, *MOON2.0* consistently retrieves Rank 10 product for corresponding image query, preserving salient descriptors including "Cream-colored Bear" and "Plush". Moreover, at Rank 50, the retrieved item remains semantically coherent with the core concept "Bear", demonstrating *MOON2.0* 's robust capability to maintain relevance across increasingly broad candidate sets.

Further validation is provided in Figs. 9 and 10, which illustrate the performance of *MOON2.0* in $q^{i} \rightarrow c^{mm}$ and $q^{mm} \rightarrow c^{mm}$ scenarios, respectively. These results corroborate the model's superior capacity for multimodal feature extraction and semantic alignment. Notably, in the $q^{mm} \rightarrow c^{mm}$ setting, *MOON2.0* effectively fuses textual and visual cues from the query and can successfully identify key semantic subjects in relatively cluttered images, yielding retrieval results that are not only highly relevant but also exhibit strong semantic consistency.

In summary, these qualitative analyses collectively demonstrate that *MOON2.0* consistently retrieves products whose visual and textual representations accurately reflect the underlying intent of multimodal queries, thereby delivering search outcomes that are both highly relevant and semantically well-aligned.

## 10. Hyperparameter sensitivity

Performance is stable across key hyperparameters: varying $\alpha, \beta \in \{0.1, 1, 10\}$ or $\tau \in \{0.05, 0.1, 0.2\}$ causes $<2\%$ fluctuation in R@1 on *MBE2.0* validation set.

## 11. Effectiveness of Modality-driven MoE

To validate the effectiveness of the proposed Modality-driven MoE, we first examine its routing dynamics. During early training, the expert entropy decreases by 62%, and the routing weights $\omega_m$ stabilize for each specific objective,


Figure 8. $q^{\mathrm{t}} \rightarrow c^{\mathrm{mm}}$ retrieval cases, including three text queries and their retrieval results.


Figure 9. $q^{\mathrm{i}} \rightarrow c^{\mathrm{mm}}$ retrieval cases, including three image queries and their retrieval results.

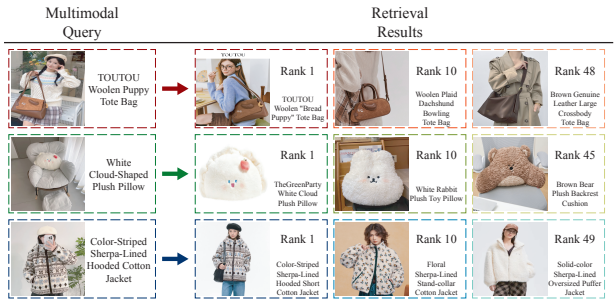


Figure 10. $q^{\mathrm{mm}} \rightarrow c^{\mathrm{mm}}$ retrieval cases, including three multimodal queries and their retrieval results.

strongly confirming the emergence of expert specialization.

Furthermore, this architecture demonstrates remarkable robustness against unbalanced data distributions. Under a text-dominant training ratio (1:10:1), *MOON2.0* maintains a competitive R@1 of 39.92 on $q^{\mathrm{i}} \rightarrow c^{\mathrm{mm}}$. In stark contrast, a tuned baseline utilizing a standard MoE with a fixed ratio suffers a severe performance degradation, dropping to 18.73 (Table 6).

Finally, regarding the ablation study, the variant denoted as "w/o Modality-driven MoE" concurrently removes both the modality-aware routing matrix $W^*$ and the sparsity loss. Further granular ablation reveals that omitting $W^*$ alone leads to a substantial decrease of 9.2 in R@1, while reverting to a conventional token-level MoE results in a 4.67 drop in R@1 on the $q^{\mathrm{i}} \rightarrow c^{\mathrm{mm}}$ retrieval task.

Table 6. Robustness to training modality ratios ($q^{\mathrm{i}} \rightarrow c^{\mathrm{mm}}$ R@1).

| Ratio (Img:Text:MM) | *MOON2.0* | Base (w/o $W^*$) |
|---|---|---|
| 12:3:2 (MOON) | 41.07 | 32.18 |
| 1:1:1 | 40.89 | 28.45 |
| 1:10:1 | 39.92 | 18.73 |

# 12. Co-augmentation Detail

## 12.1. Component Cost Analysis

Applying co-augmentation to 5 million samples requires approximately 6 days, which encompasses 85 hours of image generation across 256 A100 GPUs, alongside subsequent post-processing. This strategy is employed exclusively during the offline model training phase, yielding an absolute improvement of >6% in R@10 on retrieval tasks.

## 12.2. Prompts

To further clarify the implementation of our visual expansion procedure, we provide the exact prompts used to generate multi-granularity visual variations. All prompts were constructed and executed within a Chinese-language prompting environment, and slight deviations may occur when directly applying them in English-based settings.

## 12.3. Showcases

To complement the discussion in the main paper, we present additional co-augmentation cases illustrating how the proposed *MLLM-based Image–text Co-augmentation* is applied in practice. Each case includes the original product title and image, followed by the enriched title and several forms of visual expansion, covering main subject image extraction, background diversification, angular variation, and refinements of logos and fine-grained details. Representative examples are shown in Fig. 11.

Specifically, main subject image extraction was performed using the prompt in Algorithm 1; angular variation used the prompt in Algorithm 2; logo refinement relied on the prompt in Algorithm 3; and detail enhancement employed the prompt in Algorithm 4. For background enrichment, the initial background references and its prompt generation are given in Algorithm 5 and Algorithm 6, respectively, while the final background diversification used the prompt in Algorithm 7.

# 13. Limitation

We trained and validated our model on e-commerce data. As the field advances, there is an increasing demand

for more robust and generalizable models and methods in the field of general-purpose multimodal representation. Modality-balanced methods in this area remain underexplored and warrant further investigation, thereby establishing a clear direction for our future research.

### Algorithm 1 : Main Subject Image Extraction

**Product Description**: <Product Description>
**Product Image**: <Product Image>
**Task Instruction**: Edit the product image based on the product description and the given image, retaining only the product mentioned in the description and making it the primary focus.
**Requirements**:
1. Retain only the product described in the description; exclude any people, body parts (e.g., head, arms, legs), or other irrelevant objects.
2. Preserve the product's original appearance, including its design, color, material, and style, without alteration.
3. If any part of the product is cropped, occluded, or missing, reasonably complete it based on the image structure and product characteristics to ensure a natural, proportional, and realistic appearance.
4. Remove all irrelevant background elements, watermarks, text, and any non-product components to maintain a clean and uncluttered output; however, product logos must be preserved.
5. Standardize the background to either pure white or transparent, ensuring no shadows, reflections, or color noise.
6. Avoid illustration-like, rendered, or stylized effects; maintain a realistic photographic quality.
7. The final output should clearly present the product structure and be suitable for visual retrieval and e-commerce listing scenarios.

### Algorithm 2 : Angular Variation

**Product Image**: <Main Subject Image>
**Task Instruction**: Generate an image of the product from a different shooting angle while ensuring consistency with the product shown in the original image.
**Requirements**:
1. Maintain consistency in the product's primary structure, color, and material to ensure it remains identifiable as the same item.
2. The background may be automatically generated according to the product category, forming an e-commerce–appropriate display setting, but it should still keep the product as the main focus.
3. The final image should be clear, natural, and aligned with e-commerce listing standards.

### Algorithm 3 : Refinements of Logos

**Product Description**: <Product Description>
**Product Image**: <Main Subject Image>
**Task Instruction**: Based on the product description and image, generate an e-commerce display image that enhances the product's brand identity or brand style.
**Requirements**:
1. Preserve the product's authenticity, including its structure, color, material, and appearance, while enhancing brand-specific visual features. If a clear logo or brand element is present, emphasize, enlarge, or refine it to improve recognition. If no visible logo exists, extend existing patterns, color schemes, or textures to organically integrate brand-consistent elements into the background or composition, without introducing new logos.
2. The composition may incorporate a brand-consistent accent or emblem in the upper-left or upper-right corner to reinforce brand impression.
3. The background should be automatically expanded based on the product content or logo characteristics, adopting a clean, soft-light, e-commerce–appropriate style that maintains visual coherence and high quality.
4. Do not alter the product's design, structure, or material.
5. The final output should meet e-commerce display standards: soft lighting, clear product presentation, noticeable brand characteristics, and an overall commercial-grade aesthetic.

### Algorithm 4 : Refinements of Fine-grained Details

**Product Description**: <Product Description>.
**Product Image**: <Main Subject Image>.
**Task Instruction**: Based on the product description and the original product image, generate a close-up display image that highlights key representative details while maintaining full consistency with the overall product.
**Requirements**:
1. Product integrity: The product must remain entirely consistent with the original image; no part of the product may be regenerated, altered, or modified.
2. Detail identification: Automatically identify visually important or representative detail regions of the product and present them through a proportionally appropriate close-up enlargement.
3. Natural presentation: The enlarged region should transition naturally from the overall composition, using close-up or cropped-detail views while maintaining stylistic consistency with the full product.
4. Background: Keep the background clean and soft; it may be extended or subtly blurred as needed, adhering to e-commerce display standards.
5. Commercial quality: The final output must exhibit a professional commercial presentation—clear and natural detail emphasis, proper visual hierarchy, and a realistic, artifact-free close-up view.

### Algorithm 5 : Background References

**Background References**: <Background References>
**Role**: You are a professional e-commerce visual designer with deep familiarity of display styles on major Chinese platforms.
**Instruction**: The images show typical apparel display backgrounds used on e-commerce platforms. Carefully observe the lighting, color tone, composition, and overall atmosphere of these backgrounds. For subsequent tasks, generate background prompts that match the visual style and mood of the provided examples.

## Algorithm 6 : Background Prompt Generation

**Product Description**: <Product Description>
**Product Image**: <Main Subject Image>
**Task Instruction**: Based on the product description and the given product image, generate ten background prompts that comply with e-commerce display standards.
**Requirements**:

1. The background should remain supportive and must not overshadow the product, ensuring that the item remains the primary visual focus.
2. The background must be simple, clean, realistic, and bright, avoiding complex elements or compositions that occupy a large portion of the frame.
3. A model may be included, provided that the proportions appear natural and the wearable item remains the clear visual focus.
4. Each prompt should be a single sentence, written in natural style, non-repetitive, and conveying a professional e-commerce aesthetic.

## Algorithm 7 : Background Diversification

**Product Description**: <Product Description>
**Product Image**: <Main Subject Image>
**Task Instruction**: Based on the product description and the given product image, generate a high-quality e-commerce display image consistent with mainstream platform styles.
**Requirements**:

1. Preserve the main subject: The product must remain the central focus of the image, with its appearance, color, texture, and shape kept fully consistent with the original. The product must not be altered or obscured in any way.
2. Background: <Background Description>
3. Natural background integration: The background and the product should match naturally in lighting, perspective, and color tone, avoiding any visible seams, compositing artifacts, or stylistic discontinuities.
4. Visual style: The overall image must be realistic, clear, and natural, avoiding AI-artifacts, illustration-like styles, or rendered/CGI textures.
5. Output objective: Produce a high-quality product display image suitable for e-commerce homepages, product detail pages, or brand presentation pages.


Figure 11. Additional co-augmentation cases, including enriched titles and visual expansions (main subject image, background, angular views, logos, and details).