A Dialogue-based Multimodal Retrieval Framework for Vietnamese E-commerce RAG System

1st Ho Ngoc Tuong Vy
University of Information Technology, VNU-HCM
Ho Chi Minh City, Vietnam
21521685@gm.uit.edu.vn

2nd Ngo Thuan Phat
University of Information Technology, VNU-HCM
Ho Chi Minh City, Vietnam
21522445@gm.uit.edu.vn

3rd Nguyen Huynh Minh Huy
University of Information Technology, VNU-HCM
Ho Chi Minh City, Vietnam
huynhm.19@grad.uit.edu.vn

4th Nguyen Minh Nhut
University of Information Technology, VNU-HCM
Ho Chi Minh City, Vietnam
nhutnm.17@grad.uit.edu.vn

5th Nguyen Dinh Thuan
University of Information Technology, VNU-HCM
Ho Chi Minh City, Vietnam
thuannd@uit.edu.vn

***Abstract*—Building robust retrieval for Vietnamese e-commerce requires handling multimodal data—images, text, and conversational edits—under strict latency constraints, while recent studies highlight the promise of dialog-based retrieval for interactive product search. Existing systems still rely on multi-stage pipelines with costly reranking. We present a unified retrieval framework with three tightly integrated modules: (1) an Attribute Predictor that captures stable product semantics, (2) a Contrastive Product Captioner that generates concise comparative feedback between items, and (3) a Dialog-based Retriever with contextualized late interaction for token-level multimodal matching. Our design performs a single query-side pass, avoiding redundant computation while enabling multi-turn conversational refinement. Evaluated on a curated womenswear corpus from Tiki, the system achieves strong attribute prediction accuracy, competitive caption quality, and improved offline and interactive retrieval performance (higher MRR/Recall@k) under tight latency constraints. To our knowledge, this is the first dialog-based, late-interaction retriever for Vietnamese e-commerce, and we will publicly release data and code to support reproducible research.**

***Index Terms*—Multimodal retrieval, Dialog-based product search, Contextualized late interaction, Attribute prediction, Contrastive captioning, Vietnamese e-commerce**

I. Introduction

Multimodal retrieval is increasingly critical for e-commerce, where product search must integrate visual appearance, textual attributes, and user preferences across diverse modalities. While recent work on multimodal fusion and dialog-based retrieval shows promise [2], [3], practical deployment faces persistent challenges in efficiency, robustness, and data complexity.

Attribute-rich queries often hinge on fine-grained cues, like subtle color shades or sleeve lengths, that heuristic fusion (e.g., average, max, reciprocal-rank) tends to weaken. Multi-stage pipelines with query clarification and cross-encoder reranking improve accuracy but incur high latency and computational costs.

Vietnamese e-commerce data amplifies these issues: code-mixed text, missing diacritics, inconsistent taxonomies, and attributes shifting between text and image complicate retrieval. Across dialog turns, the dominant modality may change, sometimes text, sometimes visual cues—making fixed-weight fusion brittle and heavy reranking impractical for real-time use.

We address these challenges with a contextualized, late-interaction, dialog-based retriever that adaptively reweights token-level evidence by modality and turn, removing cross-encoder reranking. Our system combines an EfficientNet-based attribute predictor [4], a contrastive product captioner for comparative feedback [1], [2], and a ColBERT-style retriever with modality-aware token selection [5], [6]. A single query-side pass preserves token-level nuance, leverages dialog history, and improves MRR and Recall@1/5 under realistic latency constraints.

Our contributions are: (i) a unified architecture combining conversational feedback with modality-wise late interaction; (ii) an efficient pipeline removing costly reranking while retaining fine-grained matching; and (iii) empirical results on Vietnamese e-commerce showing superior performance over heuristic and single-vector baselines, pointing toward multilingual, production-ready retrieval.

II. Literature Review

Multimodal retrieval. Modern multimodal retrievers align diverse inputs—images, text, audio—into a shared representation space, typically via large-scale contrastive pretraining. Vision–language dual encoders such as CLIP and ALIGN showed the power of image–text alignment, inspiring extensions to additional modalities (e.g., ImageBind) and multilingual settings (e.g.,

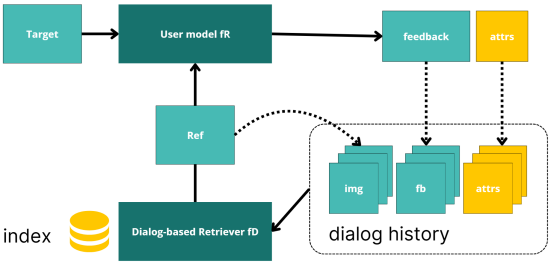


Figure 1: Dialog-based High-level Architecture

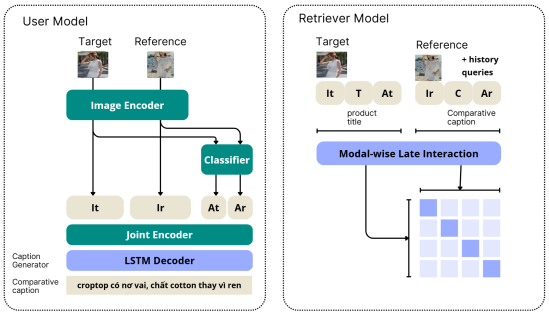


Figure 2: User and Retriever models

PaLI, VLMo) [7]–[11]. More recent work incorporates structured cues like OCR or ASR, but often relies on fixed fusion schemes that fail when only part of the input carries critical information. This motivates adaptive methods that shift focus across modalities at query time rather than using static weights.

Contextualized Late Interaction. Late-interaction retrievers address this by preserving token-level evidence until final scoring, delaying aggregation so fine-grained cues remain intact. ColBERT introduced max-similarity over contextualized token embeddings, balancing cross-encoder accuracy with bi-encoder efficiency [5]. ColBERTv2 further improved this trade-off [12], while CLaMR extended contrastive late interaction to multimodal and video retrieval with modality-aware training, outperforming simple fusion methods [6]. These results highlight late interaction's suitability when subtle cues—e.g., color terms or object parts—must survive aggressive pooling.

Dialogue-based retrieval. Dialogue-based retrieval adds user feedback, enabling iterative refinement relative to a reference item (e.g., "like this dress but in black"). FashionIQ pioneered this paradigm with human-authored comparative captions [2], inspiring interactive retrieval research. Yet many systems still use multi-stage pipelines or rigid fusion, leaving room for single-stage retrievers that integrate feedback with visual context while preserving token-level evidence—especially critical for low-resource languages like Vietnamese.

III. Methodology

Building on limitations in multimodal and dialog-based retrieval, we propose a Vietnamese e-commerce framework that integrates attribute prediction, contrastive caption generation, and dialog-based retrieval into a single, low-latency system. Our Contextualized Late-Interaction Dialog-based Retriever extends token-level late interaction to multimodal, multi-turn search, addressing noisy, unstandardized product data while maintaining efficiency and accuracy.

A. High-level Architecture

Motivated by conversational shopping behavior—where users iteratively refine requests with short comparative feedback (e.g., "like this dress but in black")—we follow the FashionIQ setup [2]. Given a reference image and turn-by-turn feedback, the system aims to quickly retrieve the desired target item using all available signals (images, texts, predicted attributes) without costly multi-stage pipelines.

B. User Model - Comparative Feedback Generator

We simulate a shopper comparing a reference and target item, refining intent via short Vietnamese captions ($\leq$ 20 tokens) highlighting key differences (e.g., color, silhouette, sleeves). The generator reuses vision backbones and predicted attributes, fusing them into a joint context for a lightweight LSTM decoder trained with cross-entropy and a small attribute-consistency term. Seed annotations plus synthetic captions expand phrasing coverage. At inference, the model provides concise, discriminative feedback to drive turn-by-turn retrieval.

The user model reuses the attribute backbone to extract visual features and attribute logits for *reference* and *target*; both streams are projected to a shared size and *joint-encoded* by a shallow MLP over

$$[\text{ref},\ \text{tgt},\ \text{ref} - \text{tgt},\ \text{ref} \odot \text{tgt}].$$

The resulting single context vector initializes an LSTM decoder that emits a short Vietnamese edit.

We use teacher forcing with token-level cross-entropy (ignoring <pad>) plus an attribute-consistency term to anchor text to visual evidence:

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \lambda_{\text{attr}} \cdot \tfrac{1}{2}\Big(\text{BCE}(\hat{A}_r, A_r) + \text{BCE}(\hat{A}_t, A_t)\Big).$$

Then, we briefly pretrain only the decoder, then joint-train the full model with AdamW, dropout, AMP, and gradient clipping—reducing caption drift on noisy, long-tailed attributes.

C. Contextualized Late Interaction for Dialog-based Retrieval

Fine-grained retrieval (e.g., "red A-line dress" vs. "burgundy A-line with shorter sleeves") requires token-level fidelity that single-vector bi-encoders cannot provide. In our FashionIQ-style setup [2], each turn yields three cues—image patches, feedback tokens, and attribute vectors—so the retriever must dynamically attend to the most discriminative channel.

We adopt *contextualized late interaction* in the ColBERT spirit [5]: queries/documents are token sets scored via max-similarity. Following CLaMR [6], we compute modality-specific scores (image/text/attributes) and select the strongest, acting like an attention switch: text dominates for "floral pattern," image dominates for visual-only edits. This reduces off-topic interference compared to early fusion or pooled embeddings.

Model setup and training: Our retriever uses **Qwen2-VL-2B** [13] as the vision–language backbone, chosen for its balance of adoption, compactness, and compatibility with efficiency methods (LoRA, 4-bit quantization) [**?**]. Qwen2-VL supplies contextual tokens, CLIP contributes robust image/text towers, and attributes are projected into the same space. Catalog embeddings are cached offline; only queries are encoded per turn.

Two-stage training.: We first *warm up* on single-turn data with contrastive learning and in-batch negatives:

$$\mathcal{L}_{\text{warm}} = \tfrac{1}{B}\sum_{b=1}^{B} \mathrm{CE}\Big([s(q_b, p_b^+)/\tau,\ s(q_b, n_{b,1})/\tau, \ldots],\ 0\Big).$$

Then we *fine-tune* on dialog inputs, where queries accumulate comparative captions and attributes across turns $k = 1..K$:

$$\mathcal{L}_{\text{ft}} = \tfrac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} w_k\, \mathrm{CE}\Big([s(q_{b,k}, p_b^+)/\tau,\ s(q_{b,k}, n_{b,k,1})/\tau, \ldots],\ 0\Big).$$

The final loss is $\mathcal{L}_{\text{ret}} = \mathcal{L}_{\text{warm}} + \mathcal{L}_{\text{ft}}$ (stage 2 only). Warm-up yields stable alignment; multi-turn training teaches modality routing. We optimize with AdamW, linear warm-up $\rightarrow$ cosine decay, AMP, and gradient clipping.

D. Evaluation

We evaluate each module on held-out splits with fixed seeds. Retrieval is assessed in both *offline* (static query) and *online* (simulated dialog) settings.

1) Attribute Predictor: Multi-label prediction with a fixed decision threshold (0.5 by default). We report:

- **Precision / Recall / F1 (macro):** per-attribute quality of positive detection, averaged across labels.
- **Support:** number of positives per attribute to contextualize long-tail effects.

2) Product Captioner: On reference–target pairs we generate short Vietnamese comparative captions (nucleus sampling, capped length). We report:

- **BLEU-1/2/3/4:** n-gram fidelity to references.
- **ROUGE-L:** longest-common-subsequence overlap (fluency/coverage).
- **CIDEr:** TF–IDF consensus with references (salience).
- **SPICE:** semantic tuple alignment (objects/attributes/relations).

We also track the captioner's auxiliary attribute head with standard multi-label accuracy/F1 to ensure grounding.

3) Dialog-based Retriever (offline): Rank the ground-truth item in a fixed catalog using cosine similarity over learned embeddings. We report:

- **MRR:** average reciprocal rank (overall rank quality).
- **Recall@1 / Recall@5:** hit rate at small cutoffs (top-k utility).
- **nDCG:** to evaluate the retrieval quality, if top 10 products are relevant.

4) Dialog-based Retriever (online, simulated): **Protocol.** For each example, we simulate up to T turns: (i) retrieve with the current dialog state; (ii) if the target is not at rank 1, refine the query using a new comparative caption and a lightweight attribute update; (iii) repeat. Let $t^\star \in \{1, \ldots, T\} \cup \{\varnothing\}$ be the first turn at which the target reaches rank 1 (if ever). We aggregate:

$$\begin{aligned} \text{top1\_success@}\ \leq T &= \frac{1}{|\mathcal{Q}|}\sum_{q} \mathbf{1}\{\, t^\star(q) \leq T \,\}, \\ \bar{t} &= \frac{1}{|\{\, q :\ t^\star(q) \neq \varnothing \,\}|}\sum_{q:\, t^\star(q)\neq\varnothing} t^\star(q) \end{aligned} \tag{1}$$

We also plot the per-turn curve $\mathrm{top1\_success@turn}\ t$ (fraction first succeeding exactly at turn t). Retrieval uses a catalog index built from precomputed document embeddings; the index is refreshed once per epoch to track training progress. Caption metrics reuse the same tokenization at validation and test to avoid drift.

IV. Dataset

A. Data Collection

1) Corpus Construction: We built a Vietnamese womenswear corpus centered on *dresses* and extended to nearby subcategories (skirts, tops, outerwear, etc.) to increase attribute diversity and provide natural cross-category negatives. Data were crawled from public product-listing and detail pages on TIKI with polite throttling. For each product we retained identifiers/variants (`id`, `sku`, options), textual fields (title, short/long description), imagery (thumbnail and gallery), pricing, brand, and taxonomy (categories/breadcrumbs). This

schema underlies the attribute vocabulary covering category, silhouette, color, length, pattern, material, sleeve cues, and size hints.

a) Subcategories covered.: Table I lists the womenswear subcategories included. While experiments emphasize dresses, adjacent categories enrich the attribute space and yield harder negatives.

b) Raw Processing.: Crawled data were lightly normalized: Vietnamese text preserved diacritics and removed HTML/boilerplate; galleries were curated by discarding watermarked/low-quality or near-duplicate frames; and size/color variants were consolidated under a single `master_id` to stabilize product identity.

B. Data Preprocessing

We then standardized raw records (Sec. IV-A) into a compact schema and aggressively cleaned noisy text/images. The goal was to preserve stable product semantics while removing duplication and artifacts.

a) Schema.: Identifiers, cleaned textual metadata (title, short/long description, breadcrumbs), brand/origin, options, and a vetted image set are retained; transient marketing fields are dropped. The finalized schema is shown in Table II.

Table I: Targeted womenswear subcategories from TIKI.VN.

Subcategory (urlKey)	Name
dam-vay-lien	Dresses (one-piece)
chan-vay-nu	Women's skirts
quan-nu	Women's pants
ao-vest-ao-khoac-nu	Women's blazers & outerwear
ao-lien-quan-bo-trang-phuc	Jumpsuits & outfit sets
do-doi-do-gia-dinh	Couple & family outfits
thoi-trang-bau-va-sau-sinh	Maternity fashion
thoi-trang-nu-trung-nien	Women's mature fashion
do-lot-nu	Women's lingerie & underwear
trang-phuc-boi-nu	Women's swimwear
ao-so-mi-nu	Women's shirts
ao-kieu-nu	Women's blouses & tops
ao-crop-top	Crop tops
ao-thun-nu	Women's T-shirts

Table II: Retained schema after preprocessing.

Field	Description
master_id	Consolidated product identifier
sku/options	Variant-level identifiers (size, color)
title/desc	Cleaned short and long descriptions
breadcrumbs	Category taxonomy
brand/origin	Product brand and origin info
images	Curated gallery (deduplicated, watermark-free)

b) Cleaning pipeline.: Text is normalized with light stopword pruning; images are deduplicated via perceptual hashing, filtered for overlays/watermarks, and capped to representative views. This yields a text-normalized, attribute-rich, and visually curated corpus ready for attribute prediction, captioning, and multi-turn retrieval.

C. Data Preparation

1) Attribute Predictor: We aim to prepare a reliable multi-label attribute vocabulary so that the predictor can infer *size*, *color*, *category*, *brand*, and other style cues directly from product images. This provides structured features that complement visual embeddings and support both retrieval and captioning.

Starting from the cleaned corpus (Sec. IV-B), we curated attribute labels in several steps: leftmargin=1.2em

- **Direct metadata extraction:** sizes, colors, categories, and brand names were taken directly from normalized product metadata.
- **Corpus-driven mining:** cleaned product text (titles, descriptions, breadcrumbs) was scanned to identify additional attribute phrases (e.g., materials, silhouettes, details).
- **Consolidation:** overlapping or overly generic phrases were removed, synonyms were canonicalized, and fragmented size mentions were mapped to a standard set (XS–5XL, Free).

Each product was assigned a multi-hot attribute vector, which was inherited by its vetted gallery images. To reduce redundancy, we capped the number of images per product. Extremely rare labels were pruned to mitigate imbalance.

The final preparation yields a compact attribute vocabulary ($\sim$1.3k labels) and balanced image–attribute pairs. This supervision serves as the basis for training the attribute predictor, enabling it to provide consistent, semantically rich cues across products.

2) Product Captioner: We prepare training data for comparative captioning as a short, repeatable pipeline:

1) **Pair mining.** Build product pairs (I_r, I_t) by Jaccard overlap on multi-hot attributes; keep pairs with overlap $\geq$ 0.6 and at least one differing attribute. From $\sim$1.3k products we obtain $>$5k pairs.
2) **View selection.** For each pair, prioritize gallery images with suffixes `_1-_3` (canonical views), then sample additional frames; cap at 6 images per pair.
3) **Synthetic captions.** Generate 5–8 concise ($\leq$ 20 words) Vietnamese comparative captions per pair using an LLM constrained by a JSON schema over our attribute families (color, silhouette, sleeve, material, *etc.*); deduplicate and validate, retaining $\sim$4.3k high-quality captions.

4) **Triplet expansion.** Combine views and captions to form (I_r, I_t, y) triplets, yielding $\sim$145k samples.
5) **Attributes & vocab.** Attach per-image multi-hot vectors aligned to a global vocabulary ($L \approx 1355$); merge attribute surface forms into the captioner vocabulary to avoid `<unk>`.
6) **Length filter.** Keep captions in 10–20 words (matches the natural distribution and stabilizes an LSTM decoder).

3) Dialog–based Retriever: We construct compact training triplets for the dialog–based retriever as $(q, p, \mathcal{N})$:

$$\begin{aligned} q &= (I_r, \mathbf{a}_r, \texttt{text}_r, \texttt{history}), \\ p &= (I_t, \mathbf{a}_t, \texttt{text}_t), \\ \mathcal{N} &= \{(I_j, \mathbf{a}_j, \texttt{text}_j)\}_{j=1}^{K} \end{aligned} \tag{2}$$

The pipeline is kept minimal and implementation–aligned:

1) **Ref–Target pairing.** Reuse captioner pairs: pick a reference image (I_r) as the current dialog state; set the target (I_t) as the desired product with its cleaned title $\texttt{text}_t$.
2) **Attach signals.** Compute multi–hot attributes $(\mathbf{a}_r, \mathbf{a}_t)$; generate a short comparative caption $\texttt{text}_r$ from the captioner; optionally include a lightweight `history` (recent texts/ids).
3) **Negatives.** Mix in–category random negatives (same breadcrumb) with *hard* negatives from a temporary cosine index; ensure fixed K per sample (pad if needed).
4) **Normalization & caching.** Images: resize+center–crop to 224, scale to $[0, 1]$. Text: Vietnamese cleaning consistent with earlier stages. Attributes: float32 multi–hot aligned to the global vocab. IDs: normalized for stable HDF5/memmap caching.
5) **Batching.** The loader yields exactly $(\texttt{query}, \texttt{pos}, \texttt{neg\_list})$ with fixed K, ready for contrastive loss and modality–wise late interaction.

V. Result

A. Experimental Environment Setup

All experiments ran on Google Colab with GPU acceleration. We split workloads by model size: leftmargin=1.2em

- **Attribute Predictor & Product Captioner:** single NVIDIA T4 (16GB); mixed-precision (AMP) training.
- **Dialog-based Retriever:** NVIDIA A100 (40GB) for Qwen2-VL-2B; LoRA fine-tuning and optional 4-bit quantization; supports multi-turn contexts.

Stack: Python 3.10, PyTorch 2.2, HF Transformers 4.41. Cleaned product JSON and images were cached in drive and indexed via HDF5 for fast, reproducible I/O.

B. Attribute Predictor

We enrich product representations with a multi-label *attribute predictor* trained on the curated corpus (Sec. IV-B). Images pass through frozen pretrained backbones (EfficientNet-B0/B4, Swin) and a linear head. We optimize only the head with BCE-with-logits (Adam, lr 10^{-4}), standard augmentations, AMP, and early stopping (typically converging within 10–20 epochs). At inference we keep the probability vector $\hat{\mathbf{a}} = \sigma(\mathbf{s})$ (before thresholding) as stable semantic features for captioning and retrieval.

Table III: Attribute prediction performance across different backbones (macro metrics).

	Precision	**Recall**	**F1**
EfficientNet-B0 + Linear	0.936	0.888	0.909
EfficientNet-B4 + Linear	0.949	0.921	0.932
Swin Transformer + Linear	0.949	0.936	0.941

As in Tab. III and Fig. 3, Swin yields the best F1, B4 offers a strong accuracy–efficiency trade-off, and B0 is the lightest with minor degradation. All three provide reliable semantic enrichment for downstream tasks.

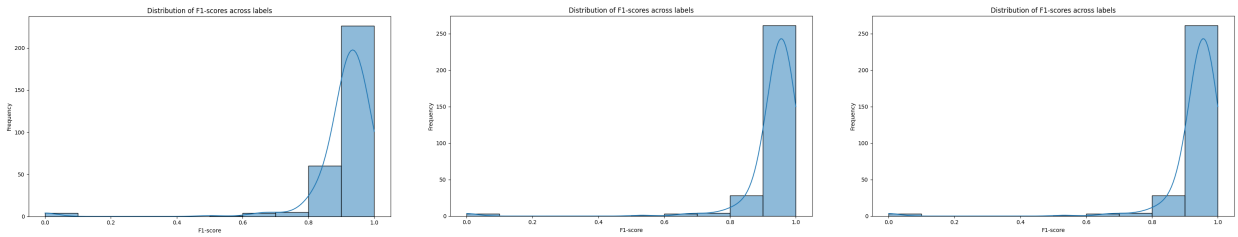
C. Product Captioner

This is a lightweight captioner to generate short Vietnamese *comparative* captions that describe how a target differs from a reference. Visual features and attribute vectors are fused into a joint context that initializes an LSTM decoder (captions $\leq$ 20 tokens). Training uses teacher forcing with cross-entropy plus an attribute-consistency regularizer, AdamW (lr 5×10^{-5}), AMP, dropout, gradient clipping, and early stopping (typically 15–20 epochs).

EfficientNet backbones (especially B4) perform best for short comparative captions; B0 is competitive and lightweight. Swin underperforms with this simple LSTM setup.

D. Modality-wise Late Interaction Dialog-based Retriever

Finally, the evaluation of *offline* (single-shot ranking) and *online* (simulated multi-turn refinement), reporting MRR/Recall@1,5 and turn-based success. For offline, we evaluate directly on the test dataset. And for online, we simulated the dialog by using the reference product to search for the target product in a vector database, then stack up the history with the used reference, and set the current reference to the top-1 product found.



(a) EfficientNet-B0 (b) EfficientNet-B4 (c) Swin Transformer

Figure 3: Attribute predictor: per-class accuracy distributions across backbones.

Table IV: Product captioning performance across backbones.

	BLEU-4	**ROUGE-L**	**CIDEr**	**SPICE**
EfficientNet-B0 + LSTM	0.316	0.566	2.859	0.185
EfficientNet-B4 + LSTM	0.331	0.583	3.008	0.195
Swin Transformer + LSTM	0.017	0.298	0.139	0.025

Table V: Dialog-based retrieval performance (Efficient Net B0 as captioner backbone).

	Offline Retrieval	**Online Multi-turn Success**
Metrics	MRR = 0.663	Dialog@$\leq$T1 = 0.35
	Recall@1 = 0.475	Dialog@$\leq$T3 = 0.45
	Recall@5 = 0.915	Dialog@$\leq$T5 = 0.65
	nDCG@10 = 0.743	Mean Turns = 3.41
Model	Contextualized Retriever	

VI. Conclusion and Further Development

We introduced a compact, end-to-end pipeline for Vietnamese e-commerce retrieval that couples structured semantics (multi-label attributes), natural feedback (comparative captions), and efficient ranking (modality-wise late interaction). The design is shaped by real marketplace constraints—noisy listings, long-tail attributes, and tight latency—and aims to turn them into advantages: attributes stabilize meaning beyond pixels; short, grounded captions expose the exact edits a shopper intends; and late interaction preserves token-level evidence without the cost of a cross-encoder.

The results suggest that lightweight vision backbones are sufficient to produce reliable attribute signals; synthetic, attribute-aware captions are a practical proxy for user feedback; and selecting the strongest modality at scoring time is crucial for fine-grained distinctions (for example, color vs. sleeve vs. texture) across dialog turns. Together, these choices yield dialog-friendly retrieval that remains accurate while meeting latency budgets.

However, the current scope emphasizes womenswear and relies partly on synthetic feedback. Next steps include broadening categories and sellers to stress long-tail robustness, adding human-in-the-loop dialogs to calibrate caption style and difficulty, exploring stronger yet compact vision–language backbones with distillation for on-device serving, and releasing the curated corpus and code to support reproducible, multilingual, multimodal product search.

VII. Acknowledgment

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2020. doi: 10.48550/arXiv.2002.05709.

[2] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. arXiv: 1905.12794.

[3] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. [Online]. Available: CVPR OpenAccess.

[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2019. arXiv: 1905.11946.

[5] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *Proc. SIGIR*, 2020. arXiv: 2004.12832.

[6] H. Lu *et al.*, "CLaMR: Contrastive Late Interaction for Multimodal Retrieval," *arXiv preprint*, 2024. arXiv: 2403.05525.

[7] A. Radford *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021. [Online]. Available: PMLR.

[8] C. Jia *et al.*, "Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision (ALIGN)," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021. [Online]. Available: PMLR.

[9] R. Girdhar *et al.*, "ImageBind: One Embedding Space to Bind Them All," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv: 2305.05665.

[10] X. Chen *et al.*, "PaLI: A Jointly-Scaled Multilingual Language-Image Model," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023. [Online]. Available: OpenReview.

[11] H. Bao *et al.*, "VLMo: Unified Vision–Language Pre-training with Mixture-of-Modality-Experts," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [Online]. Available: Proceedings.

[12] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction," in *Proc. NAACL-HLT*, 2022. [Online]. Available: ACL Anthology.

[13] S. Bai *et al.*, "Qwen2.5-VL Technical Report," *arXiv preprint*, 2025. arXiv: 2502.13923.