UMRetail: A Unified Multimodal Dataset for Hyper-Dense Shelves in Smart Retail*

Bidong Chen
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal
bidong.chen@mpu.edu.mo

Lingui Li†
School of Modern Information Industry
Guangzhou College of Commerce
Guangzhou, China
ORCID: 0009-0004-7133-868X
mddll8945@gmail.com

Rui Pedro Paiva
Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal
ORCID: 0000-0003-3215-3960
ruipedro@dei.uc.pt

Jielong Guo
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
jielong.guo@mpu.edu.mo

Jianxiu Cai
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
jianxiu.cai@mpu.edu.mo

Xu Yang
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
xuyang@mpu.edu.mo

Yapeng Wang†
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
yapengwang@mpu.edu.mo

***Abstract*— Vision-language models and multi-task learning are advancing scene understanding toward unified multimodal frameworks. However, retail datasets are fragmented: most target a single task, and heterogeneous annotation protocols and semantic granularity impede joint training, inference, and fair benchmarking. We present UMRetail, a unified multimodal dataset of real-world retail shelves with human-verified annotations. It comprises 17,697 high-resolution images covering 3,812 product types and provides instance-level segmentation masks, product detection bounding boxes, shelf-vacancy labels, and hierarchical product descriptions (short, medium, long) ranging from concise names to detailed specifications. These harmonized, cross-task labels enable integrated training and consistent evaluation for detection, segmentation, and vacancy detection. Experimental results demonstrate that UMRetail's rich data labels provide a reliable basis for rigorous evaluations: YOLOv11 Medium achieves state-of-the-art edge-device product detection (mAP 0.551, mAP$_{50}$ 0.806); UMRetail-MTArch raises image-to-text retrieval R@1 by 143.1% vs zero-shot Chinese CLIP and hits 74.55% Top-1 in zero-shot classification for 3,812 classes, which is 5.5 times that of Chinese-CLIP (13.43%) and 31 times that of CLIP (ViT-B, 2.39%). This establishes UMRetail as a research-deployment bridge for retail scene perception.**

***Index Terms*— Multimodal Dataset, Fine-grained Retail, Instance Segmentation, Dense Product Detection, Hierarchical Text Description**

I. Introduction

Driven by the integration of AI, robotics, and IoT, retail is rapidly digitizing. Among the core applications, automated shelf management involves accurate detection, recognition, and monitoring of densely packed products, and it directly affects operational efficiency and customer experience [1]. Traditional manual inspection is labor-intensive and prone to error. Emerging unmanned and automated stores use computer vision and IoT to enable real-time shelf monitoring [2]. Multimodal understanding that fuses visual data with product semantics shows promise for product identification and vacancy detection [1]. However, real-world environments pose significant challenges, including extreme density, occlusion, fine-grained category diversity, varying packaging and viewpoints, lighting variability, and multilingual packaging across regions.

Existing datasets and benchmarks only partially address these needs. General datasets do not capture the complexity of retail shelves. Retail-focused datasets such as SKU-110K [1], RPC [3], RP2K [4], and Unitail [5] improve coverage of products and scene density, but they typically emphasize a single task, have limited geographic coverage, or lack comprehensive multimodal and multilingual annotations (e.g., vacancy labels, instance-level masks, and hierarchical product descriptions). Table I contrasts the key attributes of these datasets (i.e., general and retail-focused datasets discussed earlier) and further highlights the scarcity of unified datasets that simultaneously support dense product detection, instance segmentation, shelf-vacancy detection, and hierarchical multilingual descriptions for real-world retail shelves.

Methodologically, mainstream detectors (YOLO [2], Faster R-CNN [9], RetinaNet [10]) perform well in standard settings but their accuracy degrades under extreme product density and heavy occlusion. Vacancy detection approaches based on background modeling or layout comparison are brittle when facing illumination changes and scene variations, while

* This work was supported in part by Macao Polytechnic University through Submission Control Code fca.2b66.9c65.2 under Grant RP/FCA-04/2022; and in part by FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra.

† Corresponding author

TABLE I
KEY ATTRIBUTES OF RETAIL SHELF BENCHMARKS. "BBOX" INDICATES THE TOTAL NUMBER OF BOUNDING BOXES; "MBOX" DENOTES THE AVERAGE NUMBER OF BOXES PER IMAGE

Dataset	Image	Bbox	mBox	Class	Dense	Vacancy	Retail	Mask	Text
Shelves [6]	354	13184	37.24	10	✗	✗	✓	✗	✗
SKU-110K	11762	1730996	147.17	1	✓	✗	✓	✗	✗
Locount [7]	50394	809659	16.07	140	✗	✗	✓	✗	✗
MVTec [8]	21000	72447	3.45	60	✗	✗	✓	✓	✗
RPC	30000	367935	12.26	200	✗	✗	✓	✗	✗
RP2K	10385	384311	37.1	2388	✗	✗	✓	✗	✓
Unitail-DET	12244	1777108	145	1454	✓	✗	✓	✗	✗
Unitail-OCR	-	1454	1	1454	✗	✗	✓	✗	✓
UMRetail (Ours)	17697	**2236209**	**159.6**	**3812**	✓	✓	✓	✓	✓

Note: ✓ denotes the presence of the attribute; ✗ denotes the absence

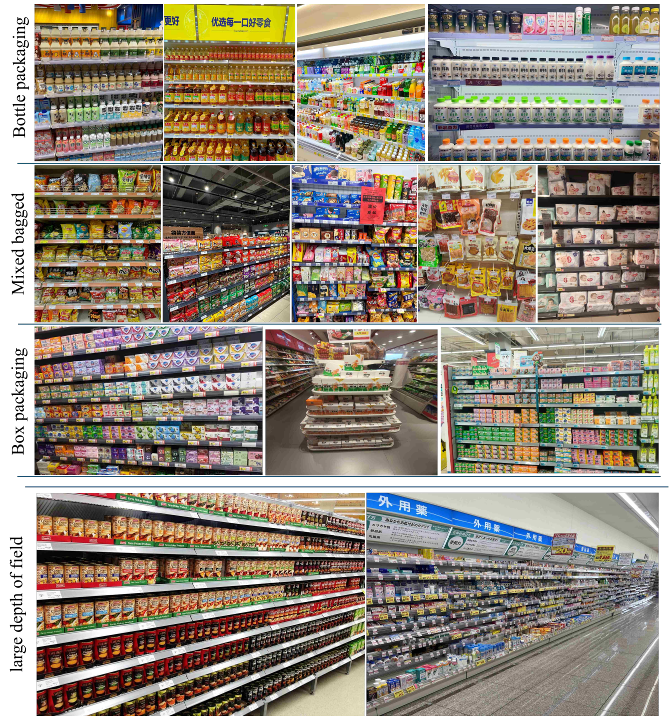


Fig. 1. Examples from the UMRetail dataset, illustrating a range of acquisition scenarios (bottle-packaged displays, mixed-bagged shelves, box-packaged racks, and large-depth of field panoramas) and annotation challenges such as vacancy regions, object scale variation, background clutter, occlusion, and overlapping products.

detector-based methods struggle with visually similar SKUs and strong occlusions. Instance segmentation in dense shelf scenes (e.g., SAM [11], Mask R-CNN [12]) is challenged by adjacent products with high visual similarity. Multimodal models (CLIP [13], FG-CLIP [14], Long-CLIP [15], Chinese-CLIP [16]) enable vision–text alignment but remain constrained by architectural limitations as well as the scarcity of fine-grained, multilingual retail annotations.

To bridge these gaps, we present UMRetail, a unified multimodal dataset of real retail shelves featuring dense layouts, diverse lighting and viewpoints, and multilingual packaging. It spans multiple geographic regions and provides harmonized annotations, including instance-level segmentation masks, product-detection boxes, shelf-vacancy labels and boxes, and hierarchical multilingual texts in short, medium, and long forms, enabling cross-lingual and fine-grained understanding. We also release benchmarks for product detection, instance segmentation, and shelf-vacancy detection across regions and languages, demonstrating UMRetail's value for research and deployment. The main contributions can be summarized as:

- **Large-scale, in-the-wild multimodal resource:** 17,697 images, 3,812 categories, 2,236,209 bounding boxes, 5,590,523 masks, and 26.8 million hierarchical multilingual descriptions.
- **Unified annotations:** instance-level masks, dense annotations of product and vacancy bounding boxes, and short/medium/long-length texts aligned with regional languages and corresponding image regions.
- **Comprehensive benchmarks:** standardized dataset splits and evaluation protocols for three core tasks, namely detection, segmentation, and shelf vacancy detection, under cross-regional and cross-lingual scenarios.

II. METHODOLOGY

A. Image Capture

The UMRetail dataset is designed to maximize the diversity, realism, and challenge of retail shelf scene understanding. Data collection spans over 100 retail stores in 30 cities across 10 countries/regions, covering a wide range of retail environments, store types, and shelf layouts. Images are captured using consumer-grade cameras and smartphones under varied lighting conditions, times of day, and operational scenarios, authentically reflecting the complexities of contemporary retail.

A distinctive feature of UMRetail is its deliberate inclusion of diverse camera perspectives. As shown in Figure 1, shelf images are captured from multiple viewpoints, including:

- Present merchandise-wall views that provide macroscopic overviews of entire shelf layouts and structure.
- Show frontal views that reveal detailed product arrangements and frontal packaging information.
- Capture oblique side views that illustrate lateral organization and product stacking.
- Depict upward and downward views that display vertical distribution and items on top or bottom shelves.

In addition, we include random placement scenarios (e.g., restocking and promotional adjustments) to increase realism and complexity.

B. Evaluation Metrics

To ensure high-quality and consistent annotations, we adopt a rigorous, reproducible protocol. We first present the evaluation metrics and formulas (for bounding boxes, segmentation masks, and text), then detail the annotation team, workflow,

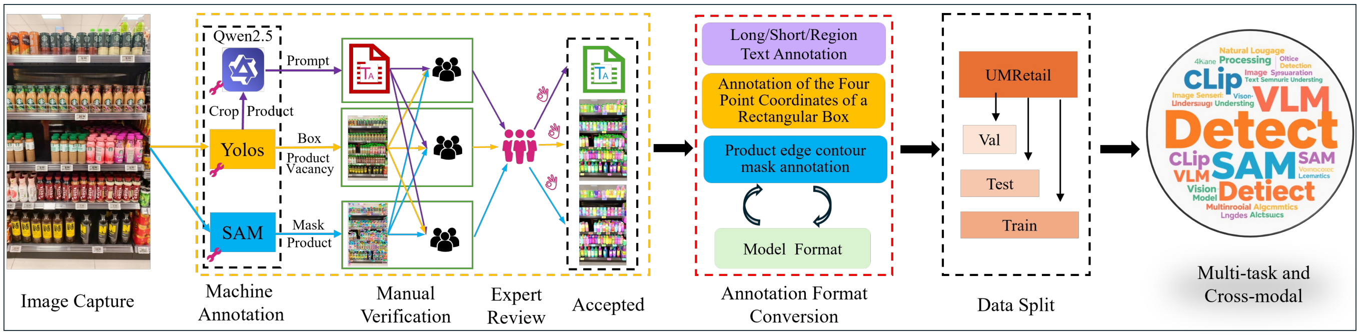


Fig. 2. UMRetail Pipeline. It illustrates the overall pipeline of dataset annotation, preprocessing, model training, and evaluation in UMRetail.

semi-automated bootstrapping procedures, and quality control, and finally present results under the defined metrics.

Audit design. We conduct three audit rounds without replacement, each sampling 3% of the dataset. Three experts independently review and correct each sampled image. For every object, we compute the Intersection-over-Union (IoU) against expert-corrected references and average across objects and rounds. Text quality is assessed via Chinese-CLIP cosine similarity and a thresholded match rate.

IoU. We first define the basic overlap metric used throughout this section. For two regions A and B in the image domain (e.g., the pixel sets of a predicted mask and a ground-truth mask),

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Mean IoU for boxes and masks. Building on the above definition, we aggregate object-level overlaps over the audited samples, where the mean IoU is computed for both bounding boxes and instance masks as follows:

$$\mathrm{mIoU}_{\mathrm{box}} = \frac{\sum_{r=1}^{3} \sum_{i \in S_r} \sum_{o \in \mathcal{O}_i} \mathrm{IoU}\big(B_{i,o}^{\mathrm{exp}}, B_{i,o}^{\mathrm{ann}}\big)}{\sum_{r=1}^{3} \sum_{i \in S_r} |\mathcal{O}_i|}, \tag{2}$$

$$\mathrm{mIoU}_{\mathrm{mask}} = \frac{\sum_{r=1}^{3} \sum_{i \in S_r} \sum_{o \in \mathcal{O}_i} \mathrm{IoU}\big(M_{i,o}^{\mathrm{exp}}, M_{i,o}^{\mathrm{ann}}\big)}{\sum_{r=1}^{3} \sum_{i \in S_r} |\mathcal{O}_i|} \tag{3}$$

where S_r denotes the sampled image set in audit round r, $\mathcal{O}_i$ the set of objects in sampled image $i \in S_r$, and for each object $o \in \mathcal{O}_i$, $B_{i,o}^{\mathrm{exp}}/B_{i,o}^{\mathrm{ann}}$ and $M_{i,o}^{\mathrm{exp}}/M_{i,o}^{\mathrm{ann}}$ denote expert-corrected vs. annotator-provided bounding boxes and masks, respectively.

Batch acceptance criterion. A batch is qualified if the corresponding mean IoU exceeds 90% (i.e., $\mathrm{mIoU}_{\mathrm{box}} > 0.90$ and $\mathrm{mIoU}_{\mathrm{mask}} > 0.90$). This criterion eschews bipartite matching and directly reflects average overlap quality across all annotated objects.

Chinese-CLIP text similarity and match (expert as gold). For instance, j, let t_j^* be the expert (gold) text and t_j the annotator text. Encode with Chinese-CLIP and L2-normalize:

$$e_j^* = \frac{E(t_j^*)}{\|E(t_j^*)\|}, \quad e_j = \frac{E(t_j)}{\|E(t_j)\|} \tag{4}$$

and define cosine similarity $s_j = \langle e_j, e_j^* \rangle \in [-1, 1]$. The continuous similarity and thresholded match rate are

$$\mathrm{Sim} = \frac{1}{N} \sum_{j=1}^{N} s_j, \quad \mathrm{Match@}\tau = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}[s_j \geq \tau] \tag{5}$$

where τ is chosen on a labeled development set (e.g., via ROC).

C. UMRetail Pipeline

The UMRetail dataset construction process is illustrated in Figure 2.

a) Team and Training: The annotation team consists of 12 full-time professionals, 9 with bachelor's degrees in computer vision and 3 with PhDs. All members undergo a two-week standardized curriculum covering product taxonomy, tool operation, quality criteria, and protocols for challenging scenes, and subsequently take a competency assessment prior to production annotation.

b) Workflow and Division of Labor: The workflow is organized into three specialized subgroups: (i) bounding-box verification and adjustment; (ii) instance-segmentation mask generation and refinement; (iii) product-attribute and hierarchical text-description labeling. The three PhD annotators provide cross-group supervision, oversee quality control, and arbitrate complex or ambiguous cases.

c) Machine Annotation and Manual Verification: In the initial phase, we adopt a semi-automated strategy: bounding-box proposals are generated using a pretrained YOLOv11 model, segmentation masks are produced via the SAM model, and multi-granular textual descriptions are auto-generated by Qwen2.5-VL [17]. These outputs serve as efficient references for subsequent manual refinement.

d) Attributes and Text Labels: For product attribute and text label annotation, we use a "machine annotation + human verification" two-stage workflow. For each shelf image, Qwen2.5-VL produces structured product descriptions, including product location, stock status, category, product name, brand, packaging color, and place of origin. Annotators cross-check all fields against the original images to ensure correctness, consistency, and completeness. Each product is annotated with three levels of text: short name, mid-level attributes, and detailed specifications, providing rich semantic context for downstream tasks.

e) Expert Review: Each subgroup first optimizes its assigned annotations, followed by cross-group verification and collective reviews to ensure global consistency and accuracy. We implement PhD-led arbitration for disagreements, 10% random-sampling audits, and weekly review meetings to iteratively refine the annotation guidelines and protocols.

f) Accepted: Under the above evaluation, we obtain:

- Bounding boxes: mean IoU of 95.2%.
- Instance segmentation masks: mean IoU of 92.6%.
- Product attributes and textual descriptions: categories, brands, packaging, volumes, and flavors are comprehensively annotated; each item contains three textual levels and undergoes the "machine annotation + human verification" process.
- Text consistency: 91.8%, measured as field-level accuracy against expert-corrected references.

g) Data application: Following annotation verification, the pipeline standardizes diverse annotations (textual descriptions, bounding-box coordinates, and contour masks) into a model-compatible format during the annotation format conversion stage. The formatted dataset is partitioned into training, validation, and test subsets via data split, enabling iterative model optimization and unbiased performance evaluation. Finally, the multi-task and cross-modal module leverages frameworks like CLIP (for vision-language encoding) and SAM (for segmentation) to construct unified models capable of simultaneously addressing detection, segmentation, and semantic understanding tasks in retail scenarios.

D. UMRetail Dataset Visualization

As shown in Figure 3, the UMAP [5] visualization of the UMRetail dataset (color-coded by meta-categories and split into train/val/test) demonstrates globally consistent clustering structures across splits, indicating that the partition preserves intrinsic feature manifolds and supports unbiased generalization assessment. Within categories, samples form compact clusters (e.g., food), while distinct categories are clearly separated, validating intra-class coherence and inter-class separability. The dataset also exhibits a long-tailed distribution (e.g., sparse outdoor products vs. dominant food), which reflects the authentic distribution of the UMRetail dataset, where inter-class imbalance itself constitutes a key challenge in intelligent retail scenarios. Minor variations across splits mainly arise from natural data collection factors in retail environments, such as differences in viewing angles (front, back, side), brand appearances, and packaging styles. These variations do not represent noise but rather faithfully mirror the inherent distribution characteristics of real-world retail data.

E. Single-task and Cross-modal Benchmarks

We conduct benchmarks on UMRetail across four core tasks:

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT-SCALE MODELS ON JETSON EDGE COMPUTING PLATFORM

Models	Scales	Precision	Recall	mAP	mAP_{50}	F1	Params	GFLOPs
YOLOv5	nano	0.781	0.710	0.514	0.778	0.744	2.50	7.1
YOLOv6	nano	0.781	0.710	0.514	0.777	0.744	4.23	11.8
YOLOv10	nano	0.781	0.705	0.514	0.776	0.741	2.70	8.2
YOLOv11	nano	0.782	0.713	0.516	0.779	0.746	2.58	**6.3**
YOLOv9	tiny	0.785	0.712	0.520	0.782	0.747	**1.76**	6.4
YOLOv8	**nano**	**0.787**	**0.714**	**0.520**	**0.782**	**0.749**	3.01	8.1
YOLOv8	small	0.785	0.719	0.519	0.783	0.751	9.82	23.3
YOLOv10	small	0.787	0.724	0.533	0.790	0.754	8.03	24.4
YOLOv6	small	0.789	0.732	0.531	0.793	0.759	15.98	42.8
YOLOv5	small	0.792	0.730	0.534	0.794	0.760	7.82	**18.7**
YOLOv11	small	0.792	0.730	0.534	0.795	0.760	9.41	21.3
YOLOv9	**small**	**0.792**	**0.738**	**0.542**	**0.800**	**0.764**	**6.19**	22.1
YOLOv6	medium	0.788	0.741	0.542	0.799	0.764	51.25	158.3
YOLOv9	medium	0.789	0.737	0.536	0.796	0.762	20.01	76.5
YOLOv10	medium	0.790	0.737	0.544	0.798	0.763	**16.48**	63.4
YOLOv8	medium	0.791	0.744	0.546	0.803	0.767	23.22	67.4
YOLOv5	medium	0.794	0.743	0.546	0.804	0.768	22.13	**52.5**
YOLOv11	**medium**	**0.797**	**0.744**	**0.551**	**0.806**	**0.770**	20.03	67.7
EfficientDet [18]	large	0.256	0.289	0.256	0.427	0.2715	**3.83**	**3.60**
RetinaNet [10]	large	0.273	0.305	0.273	0.438	0.2881	74.14	81.74
Faster-RCNN [9]	large	0.275	0.307	0.275	**0.445**	0.2901	41.35	90.9
DenseDet [19]	large	**0.313**	**0.486**	**0.313**	0.377	**0.3807**	74.14	162

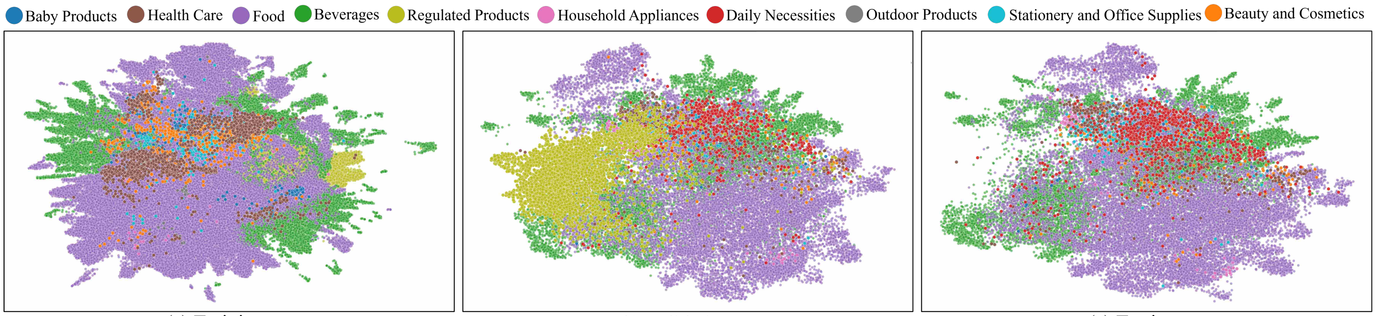
- **Product detection:** to localize and classify all product instances for inventory management, planogram compliance, and retail analytics.
- **Vacancy detection:** to identify empty shelf slots for vacancy monitoring and automated replenishment.
- **Vision–language alignment (CLIP):** to evaluate text–image alignment through retrieval and zero-shot or linear-probing recognition using multi-level descriptions and attributes.

In addition, UMRetail supports tasks such as multi-label classification, fine-grained recognition, cross-modal retrieval, and language-guided segmentation, thereby providing a comprehensive platform for the development and evaluation of vision and vision–language methods in realistic retail scenarios.

III. EXPERIMENTS

A. Object Detection Benchmark

We benchmark single-stage and two-stage detectors, including YOLOv5–11 (nano/small/medium), RT-DETR, and Mask R-CNN, initialized from COCO and trained on UMRetail under a unified protocol. We report COCO-style AP (mAP, mAP_{50}), precision/recall/F1 computed jointly for product and vacancy detection, and GFLOPs, all evaluated on a Jetson Orin Nano (4GB). As shown in Table II, YOLO models consistently outperform classic two-stage and anchor-based baselines, including RetinaNet, Faster R-CNN, and DenseDet on hyper-dense shelves. YOLOv11-medium achieves the best overall accuracy (mAP 0.551, mAP_{50} 0.806, F1 0.770). At the small scale, YOLOv9-small provides the best accuracy–efficiency trade-off (mAP 0.542, mAP_{50} 0.800, F1 0.764) with 6.19 M parameters and 22.1 GFLOPs. In the nano tier, YOLOv8-nano attains the highest accuracy (mAP 0.520, mAP_{50} 0.782, F1 0.749), while YOLOv11-nano balances accuracy (mAP 0.516) and compute (6.3 GFLOPs). These results demonstrate steady gains across YOLO generations and underscore the challenges of dense layouts and vacancy detection on resource-constrained edge hardware.



(a) Training set (b) Validation set (c) Testing set

Fig. 3. UMAP visualization after dimensionality reduction of samples from the UMRetail dataset (samples are color-coded by meta-category). (a), (b), and (c) correspond to the training, validation, and test sets, respectively, and illustrate the cluster distribution patterns of meta-categories across different splits, supporting analysis of distribution consistency and model generalization.

B. Vision-Language Benchmarks (UMRetail-MTArch)

UMRetail-MTArch (UMRetail Multi-Task Architecture), as shown in Figure 4, is a multimodal pre-trained framework built upon the Chinese-CLIP architecture. It fine-tunes Chinese-CLIP weights and integrates object detection and instance segmentation. The model extracts global image features and local region features. It fuses three types of Chinese–English bilingual text: long captions, short labels, and regional descriptions. This design constructs multi-granularity semantics. The framework uses visual and text encoders to map cross-modal features into a shared embedding space. It adopts the Chinese CLIP loss, namely the InfoNCE loss with bidirectional image–text contrast. After computing the image–text similarity matrix and scaling it with a learnable temperature τ, the model applies a bidirectional cross-entropy objective to optimize alignment. This procedure enhances the association between Chinese characters and visual features. The model achieves a strong correlation between images and text, including fine-grained information, and it supports zero-shot prediction. UMRetail-MTArch Loss (InfoNCE, bidirectional). Let N be the batch size. Let $\mathbf{z}_i \in \mathbb{R}^d$ and $\mathbf{t}_j \in \mathbb{R}^d$ denote the ℓ_2-normalized image and text embeddings, respectively, and let $\tau > 0$ be a learnable temperature. Define the similarity as the dot product $s_{ij} = \mathbf{z}_i^\top \mathbf{t}_j$. The symmetric InfoNCE objective is

$$\mathcal{L}_{\text{i}\to\text{t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(s_{ii}/\tau\right)}{\sum_{j=1}^{N} \exp\left(s_{ij}/\tau\right)} \tag{6}$$

$$\mathcal{L}_{\text{t}\to\text{i}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(s_{ii}/\tau\right)}{\sum_{j=1}^{N} \exp\left(s_{ji}/\tau\right)}. \tag{7}$$

The final training objective is

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{i}\to\text{t}} + \mathcal{L}_{\text{t}\to\text{i}}\right) \tag{8}$$

Table III shows the cross-modal retrieval performance of Chinese-CLIP (zero-shot) and UMRetail-MTArch (Chinese-CLIP fine-tuned on UMRetail data). In the zero-shot setting, Chinese-CLIP achieves 25.4% R@1 and 45.1% mRecall (average of R@1, R@5, and R@10) for text-to-image retrieval, and 25.5% R@1 and 46.8% mRecall for image-to-text retrieval. After fine-tuning, substantial absolute improvements are observed: text-to-image R@1 increases by 24.5 percentage points (49.9% vs. 25.4%), mRecall by 26.6 points (71.7% vs. 45.1%); image-to-text R@1 improves by 36.5 points (62.0% vs. 25.5%), and mRecall by 33.9 points (80.7% vs. 46.8%). These results demonstrate the significant performance gains brought by domain-specific fine-tuning on retail data.

TABLE III
CROSS-MODAL RETRIEVAL PERFORMANCE OF CHINESE-CLIP (ZERO-SHOT) VS. UMRETAIL-MTARCH (FINE-TUNED ON UMRETAIL), DEMONSTRATING ENHANCED DOMAIN ADAPTABILITY IN RETAIL SCENARIOS

Setups	Text-to-Image(%)				Image-to-Text(%)			
Metrics	R@1	R@5	R@10	mRecall	R@1	R@5	R@10	mRecall
Chinese-CLIP	25.4	49.7	60.3	45.1	25.5	52.5	62.2	46.8
UMRetail-MTArc (Ours)	**49.9**	**78.7**	**86.3**	**71.7**	**62.0**	**87.2**	**92.9**	**80.7**

R@K: Recall at Top-K; mRecall: mean of R@1, R@5, and R@10.

TABLE IV
ZERO-SHOT PRODUCT CATEGORY CLASSIFICATION PERFORMANCE OF CLIP MODELS ON THE UMRETAIL TEST SET (3,812 CLASSES)

Model	**Accuracy(%)**		**Averages(%)**		
	Top-1	Top-5	Precision	Recall	F1-Score
CLIP(ViT-B)	2.39	6.30	3.66	2.44	2.09
long-CLIP	2.87	7.88	4.64	2.93	2.48
FG-CLIP	4.28	5.11	3.89	4.23	3.94
Chinese-CLIP	13.43	32.57	21.66	13.44	12.17
UMRetail-MTArch(Ours)	**74.55**	**92.48**	**82.8**	**74.5**	**75.0**

Table IV presents the results of the zero-shot product classification task on UMRetail. Compared with Chinese-CLIP, UMRetail-MTArch (Ours) achieves 5.5 times higher Top-1 accuracy (74.55% vs. 13.43%), 2.8 times higher Top-5 accuracy (92.48% vs. 32.57%), and 6.2 times higher F1-score (75.0% vs. 12.17%). Compared to general models like CLIP-B, its lead in Top-1 accuracy surpasses 31 times (74.55% vs. 2.39%). This order-of-magnitude performance

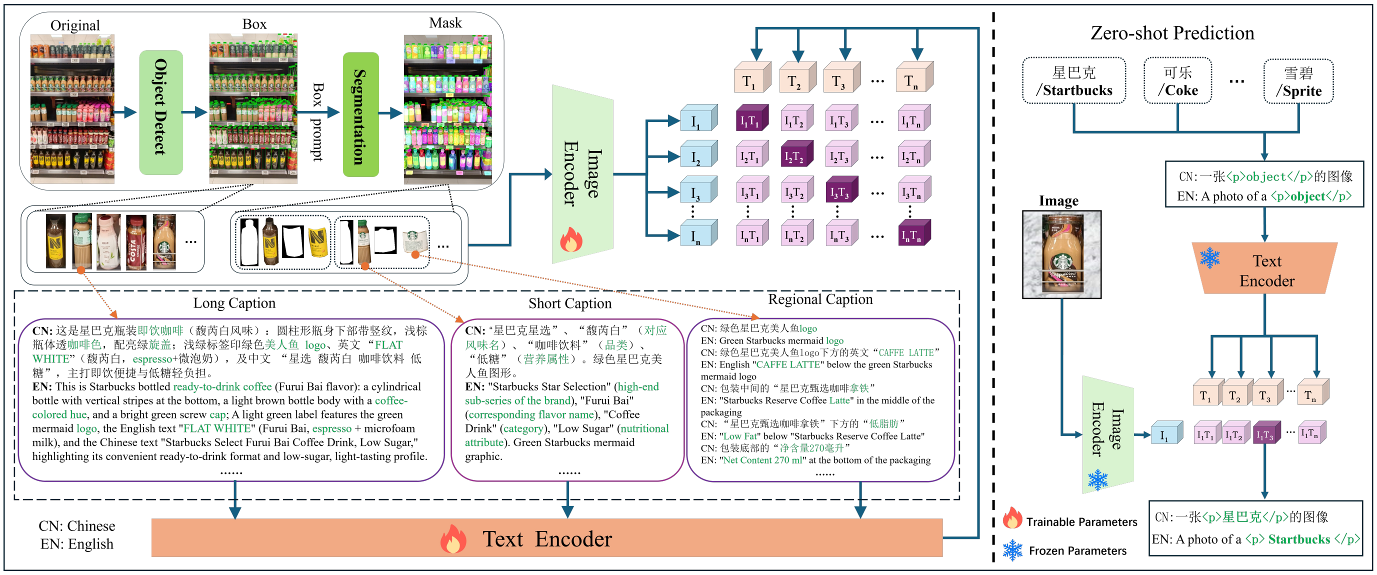


Fig. 4. UMRetail-MTArch. It extracts global-local image features via detection-segmentation, fuses multi-granularity Chinese-English texts, maps them to a shared space with Chinese-CLIP fine-tuned cross-modal encoders, aligns via contrastive loss, and enables zero-shot prediction.

gap demonstrates that domain-specific CLIP, through scene semantic adaptation, can overcome the zero-shot classification limitations of general models, providing an efficient approach for fine-grained category recognition in retail.

IV. Conclusions

Contributions: UMRetail addresses data fragmentation issues through a unified multimodal solution: it includes 17,697 high-resolution shelf images covering 3,812 product types, and provides instance masks, detection bounding boxes, empty slot labels, and hierarchical text information. Its unified and rich data annotations are crucial for cross-task implementation, enabling joint training under dense layouts and consistent cross-task evaluation. Experimental results demonstrate that UMRetail performs exceptionally well on core tasks, providing a reliable reference for edge device deployment.

Limitations: Primarily supermarket shelves with limited regional and brand diversity; mostly static images and vacancy definitions tied to specific policies; evaluations focus on Jetson Nano and omit modalities such as depth, audio, and RFID, which may limit cross-domain and cross-hardware generalization and comparability.

References

[1] E. Goldman, R. Herzig, A. Nachman *et al.*, "Precise detection in densely packed scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA: IEEE, 2019, pp. 5227–5236.

[2] Ultralytics, "Yolo-ultralytics," *Ultralytics Official Website*, n.d., accessed 2025-08-20. Resource Type: Official Website for YOLO Series Models.

[3] X. S. Wei, Q. Cui, L. Yang *et al.*, "Rpc: A large-scale retail product checkout dataset," *arXiv*, 2019.

[4] J. Peng, C. Xiao, Y. Li *et al.*, "Rp2k: A large-scale retail product dataset for fine-grained image classification," *arXiv*, 2020.

[5] P. Guimarães, J. Alves-Pinto, and H. Proença, "Unitail: Unified retail visual understanding at scale," *arXiv:2305.03958*, 2023.

[6] G. Varol and R. S. Kuzu, "Toward retail product recognition on grocery shelves," in *Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, vol. 9443. SPIE, 2015, pp. 46–52.

[7] Y. Cai, L. Wen, L. Zhang, D. Du, and W. Wang, "Rethinking object detection in retail stores," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 947–954, 2021.

[8] P. Follmann, R. König, P. Härtinger, M. Klostermann, and T. Böttger, "Mvtec d2s: Densely segmented supermarket dataset," in *European Conference on Computer Vision*, 2018, pp. 581–597.

[9] D. Arora and K. Kulkarni, "Efficient shelf monitoring system using faster-rcnn," in *Proceedings of the International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*. Chennai, India: IEEE, 2024, pp. 1–6.

[10] T. Y. Lin, P. Goyal, R. Girshick *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, 2017, pp. 2980–2988.

[11] B. Chen, L. Li, H. Zhu, M. Tan, G. Liu, H. Chi, X. Yang, and Y. Wang, "Prince: Advanced classification algorithm for rice grain recognition in clustered images," *Computers and Electronics in Agriculture*, vol. 239, p. 110949, 2025.

[12] C. Zhang, Y. Liu, H. Zhao, L. Wang, Z. Chen *et al.*, "A comprehensive survey on segment anything model for vision and beyond," *arXiv:2305.08196v2*, 2023.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[14] J. Xie, Y. Liu, Z. Wang, M. Yang, and J. Liu, "Fg-clip: Fine-grained visual recognition via clip knowledge distillation," *arXiv*, 2023.

[15] Y. Zhang, C. Li, J. Zhang, P. Gao, J. Wu, Y. Liu *et al.*, "Long-clip: Extending clip's context window for visual understanding," *arXiv:2401.02957*, 2024.

[16] A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, and C. Zhou, "Chinese clip: Contrastive vision-language pretraining in chinese," *arXiv:2211.01335*, 2022.

[17] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-vl technical report," *arXiv:2502.13923*, 2025.

[18] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[19] P. R. Florence, L. Manuelli, and R. Tedrake, "Dense object nets: Learning dense visual object descriptors by and for robotic manipulation," *arXiv:1806.08756*, 2018.