WeChat

WeMM-Ecom: Advancing Multimodal Content Understanding for WeChat E-Commerce

Binhui Xie*, Jinmin Li*, Haomiao Tang*, Guangting Wang*†
Song Jin, Jianwei Zhu, Fengyun Rao‡, Jing LYU

WeChat Vision, Tencent Inc.

Abstract

We present WeMM-Ecom, a family of multimodal large language models that deliver frontier e-commerce capabilities while achieving state-of-the-art general-purpose performance at their respective scales. The family spans five sizes, from a 0.8B dense model to a 35B mixture-of-experts model with 3B active parameters. Our training pipeline begins with two-stage continued pre-training that progressively raises the proportion of e-commerce data within a diverse multimodal data mixture, followed by hybrid-thinking post-training that combines on-policy distillation and reinforcement learning to support both non-thinking and thinking modes. For evaluation, we construct a comprehensive suite of 28 e-commerce benchmarks (23 in-house and five public) covering eight business scenarios and complement it with 64 public general-purpose benchmarks. On the e-commerce suite, WeMM-Ecom-35B-A3B surpasses substantially larger open-source flagship models, including Qwen3.8-Flash-Next, HY4, and DeepSeek-V4.1-Flash-1017. On the general-purpose suite, every WeMM-Ecom model surpasses its similarly sized Qwen3.5 counterpart, with the lightweight variants (e.g., WeMM-Ecom-2B) showing clear advantages over leading lightweight models such as Qwen3.5-2B, MiniCPM-V-4.7-1B, and LFM3-VL-2B. Beyond benchmarks, WeMM-Ecom now powers hundreds of production use cases in WeChat, including catalog management, multimodal product search, and livestream commerce, serving billion-scale requests per day. We publicly release all model weights and detail our data curation, training recipe, and benchmark construction to facilitate future work on domain-specialized multimodal models.

https://huggingface.co/collections/tencent/wemm-ecom
https://github.com/Tencent/WeMM-Ecom

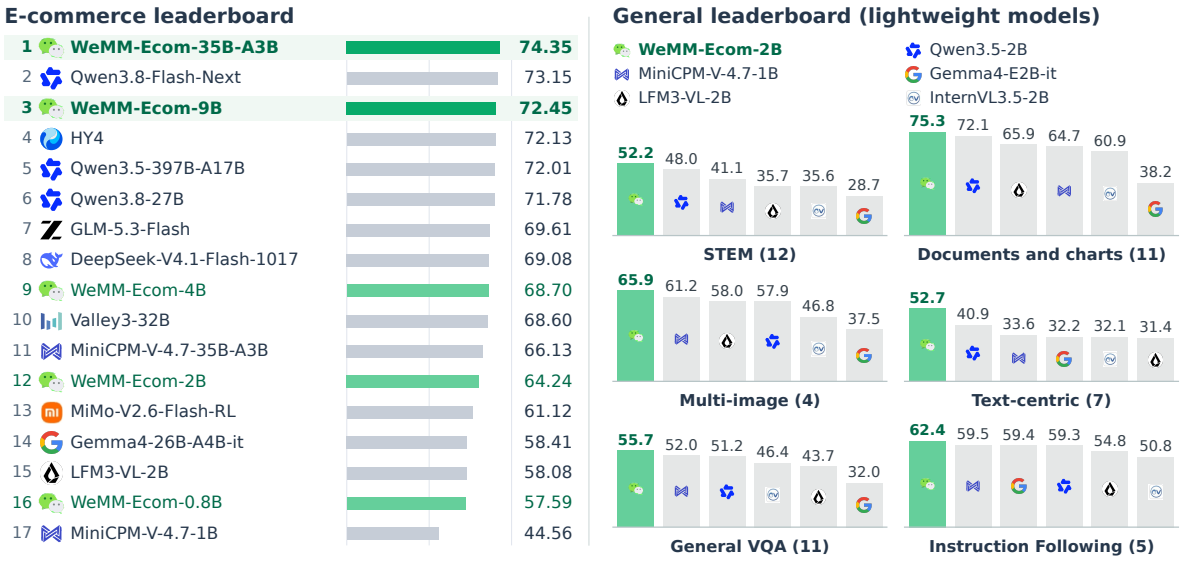


Figure 1 Comparing WeMM-Ecom against SoTA multimodal models. On the e-commerce suite, WeMM-Ecom-35B-A3B surpasses models 5×–22× larger (e.g., Qwen3.8-Flash-Next-180B and HY4-770B) (left). On the general-purpose suite, WeMM-Ecom-2B scores highest among lightweight models across six domains in non-thinking mode (right).

*Core contributors. †Project leader. ‡Corresponding author: fengyunrao@tencent.com

1 Introduction

Multimodal large language models (MLLMs) have become capable general-purpose models for understanding images, documents, and video [24, 25, 27, 69, 78, 79, 86]. In specialized domains, however, general capability alone falls short, as models must apply fine-grained domain knowledge to what they see, whether identifying a clinical finding in a medical scan [72] or extracting the correct figure from a financial chart [50].

E-commerce is an especially demanding domain for MLLMs. A product can be presented through dozens of product views, long-form detail images, and livestream frames, alongside titles and structured attribute tables. Critical evidence, such as selling points, nutrition facts, dimension annotations, and certification marks, often appears only inside images. Models must further link this visual text to the specific SKU variant in question and draw on catalog knowledge that is never explicitly stated. As e-commerce models move from text records [39, 52, 61] to multimodal understanding [7, 91], two core challenges stand out: (1) models must accurately perceive dense, interleaved visual evidence and produce outputs that satisfy strict entity-level and business-rule constraints; and (2) domain specialization must not erode general visual perception and reasoning, especially for lightweight models deployed under tight latency and memory budgets [42, 99].

To address these challenges, we present **WeMM-Ecom**, a family of multimodal foundation models spanning four dense sizes (0.8B, 2B, 4B, and 9B) and a 35B mixture-of-experts (MoE) model with 3B active parameters (35B-A3B), initialized from Qwen3.5 [69]. We curate over 40 e-commerce tasks from product images, livestream frames, and business workflows. Because these tasks directly exercise visual grounding, OCR, and multi-image reasoning, they also strengthen general perception rather than competing with it. We pair them with broad general supervision in a two-stage continued pre-training (CPT) curriculum with an increased e-commerce share in the second stage (Figure 3). Post-training then applies hybrid-thinking on-policy distillation (OPD), a single parameter-space merge with the post-trained Qwen3.5 weights that recovers general instruction-following and reasoning abilities, and group relative policy optimization (GRPO), enabling each model to support both thinking and non-thinking modes.

We evaluate WeMM-Ecom on a comprehensive 28-benchmark e-commerce suite across eight scenarios (207,117 instances) and 64 general benchmarks. Across all five sizes and both modes, WeMM-Ecom consistently outperforms size-matched Qwen3.5 on e-commerce. WeMM-Ecom-35B-A3B ranks first among evaluated flagships (Figure 1), surpassing Qwen3.8-Flash-Next, HY4, and DeepSeek-V4.1-Flash-1017. In non-thinking mode, our 0.8B and 2B models each beat a larger Qwen3.5 model on e-commerce while showing clear advantages on general-purpose capabilities. WeMM-Ecom-2B also surpasses MiniCPM-V-4.7-1B [99] and LFM3-VL-2B [42] on both e-commerce and general suites in both thinking and non-thinking modes.

Our main contributions are as follows:

1. **We open-source WeMM-Ecom**, a family of five multimodal models ranging from a 0.8B dense model to a 35B-A3B MoE model, each supporting both thinking and non-thinking modes. WeMM-Ecom delivers leading e-commerce performance at every scale while matching or surpassing its Qwen3.5 counterpart on general benchmarks, and has been validated at scale in production.
2. **We present a detailed training and data recipe for WeMM-Ecom**. We describe how more than 40 e-commerce tasks are constructed from product images, livestream frames, and business records; how they are mixed with general data and scheduled across two CPT stages; and how hybrid-thinking OPD, a one-time parameter-space merge, and GRPO are combined to produce the final models.
3. **We introduce the key design principles of our e-commerce evaluation framework**. We show how real-world business workflows are translated into eight evaluation scenarios with explicit inputs, decision targets, and scoring rules, resulting in a suite of 23 in-house and five public benchmarks with 207,117 instances.

2 Training

We initialize each WeMM-Ecom model from the corresponding Qwen3.5 Base checkpoint [69] and perform full-parameter optimization across all training stages using the WeChat-YATT framework [90]. Our pipeline first conducts CPT to instill fine-grained knowledge of products, variants, and business rules while preserving

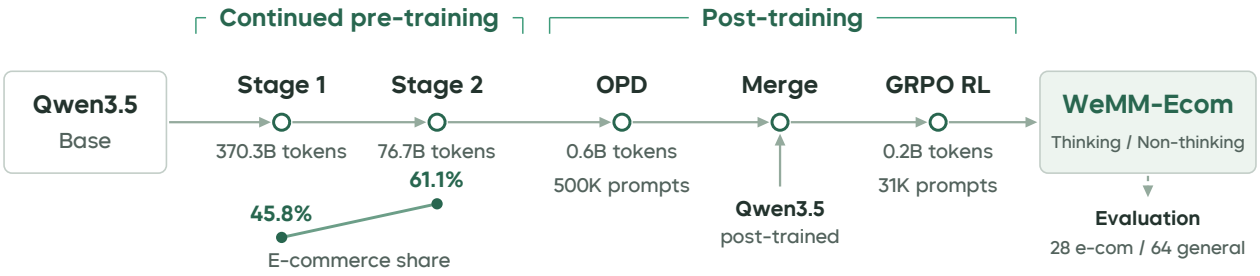


Figure 2 Training and evaluation pipeline. The e-commerce share increases in the second continued pre-training stage. Post-training combines on-policy distillation (OPD), checkpoint merging, and group relative policy optimization (GRPO). The resulting checkpoints support thinking and non-thinking modes and are evaluated on e-commerce and general benchmarks.

broad multimodal capabilities. Subsequently, post-training combines OPD for token-level teacher guidance, a single parameter-space merge with the post-trained Qwen3.5 weights, and reinforcement learning (RL) driven by task-specific verifiable rewards. All five model scales support thinking and non-thinking modes in a unified chat interface. Figure 2 summarizes the training and evaluation pipeline.

2.1 Continued Pre-Training for E-Commerce

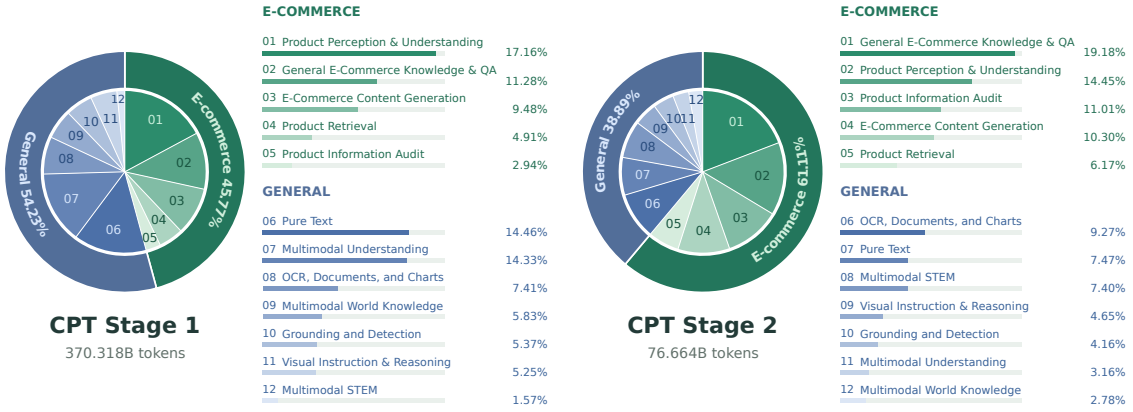


Figure 3 Token composition of the two CPT stages. Inner sectors and adjacent bars show shares across five e-commerce and seven general domains; outer rings show proportion of general and e-commerce data. E-commerce accounts for 45.8% of 370.32B tokens in Stage 1 and 61.1% of 76.66B tokens in Stage 2. Domains are ranked independently within each group and stage, thus number indices are stage- and group-specific.

CPT [43] adapts the backbone to e-commerce while preserving its general multimodal capabilities. We construct CPT data from product images, livestream frames, business records, and general multimodal and text-only corpora, and organize it into five e-commerce and seven general domains (Figure 3; Table 1). We retain both pools throughout CPT because e-commerce tasks routinely depend on general skills: reading a package label and assigning it to the requested product, for example, draws on general OCR and grounding as well as knowledge of product variants. Records with an e-commerce business context, including text-only records, are assigned to the corresponding e-commerce domain. Training proceeds in two stages; the second raises the e-commerce share and shifts emphasis from perception toward auditing and reasoning (Section 2.1.4).

Table 1 CPT domains. Each record is counted once, in the domain of its primary supervision objective rather than by source name or number of images. Records from e-commerce business sources remain in the e-commerce pool regardless of modality, and general records without visual input form Pure Text. Per-stage proportions are shown in Figure 3.

Domain	Scope
E-commerce	
Product Perception & Understanding	Category–property–value (CPV) prediction, including category prediction with reasoning, attribute extraction, and SKU specification understanding, together with localization and detection of target products in listing images, multi-image listings, and livestream frames.
Product Information Audit	Rule-conditioned review of product listings: consistency across images, titles, categories, and attributes; category mismatch; product-image quality; inappropriate content; and quality checks of short titles.
Product Retrieval	Same-product matching and retrieval at SPU or SKU granularity, including image-to-image and image-to-product search, text-based SKU matching, and similarity judgments between visual entities or product regions.
General E-Commerce Knowledge & QA	Question answering over product images and listing text, OCR of product and storefront images, shopping and price reasoning, and product, brand, and entity knowledge, including entity recognition and product-word prediction.
E-Commerce Content Generation	Product-centric text generation: per-view and overall descriptions of product images, multi-image listings, and livestream frames; short product titles; and product copy and product-sharing posts.
General	
Multimodal Understanding	Foundational understanding of single images, multiple images, and videos: dense image captions, per-image and overall multi-image descriptions, video event understanding and QA, and video temporal grounding.
OCR, Documents, and Charts	Text-rich inputs: text transcription, including multilingual OCR; document and table parsing; document QA; chart reading; and structured extraction of visible text.
Grounding and Detection	Image-space localization with coordinate targets: object detection, referring-expression grounding, GUI element grounding, and rejection of queries whose target is absent.
Visual Instruction & Reasoning	Visual tasks beyond description: VQA and cross-image comparison, counting and relational questions without coordinate targets, image-quality assessment, inferring edit instructions from source–edited image pairs, and rule-conditioned moderation of general (non-listing) content.
Multimodal STEM	Mathematics, science, and technical problems that require visual context (diagrams, charts, documents, educational material, or video), with worked solutions; includes chart-to-code tasks.
Multimodal World Knowledge	Linking visual entities and scenes to external facts: encyclopedia- and entity-grounded image descriptions, knowledge-intensive VQA, and specialized-domain tasks such as medical imaging.
Pure Text	General text-only records: mathematical and scientific reasoning, instruction-following and multilingual dialogue, encyclopedic and reading-comprehension text, and curated web text.

2.1.1 E-Commerce Data

General multimodal corpora teach a model to recognize a bag of coffee beans; e-commerce applications must also determine which variant it is—the 250 g or the 500 g SKU, a single bag or a two-pack—whether its listing complies with platform rules, and how to present it to shoppers. We therefore design e-commerce supervision around three properties that general corpora rarely provide: *variant-level granularity*, *traceable evidence*, and *rule-conditioned outputs*. The five domains in Table 1 instantiate these properties in different combinations.

Variant-level granularity. Product Perception & Understanding trains the model to resolve products at the level of categories, attributes, and SKUs. Category prediction pairs product images with taxonomy labels and reasoning traces, and attribute tasks extract visible text, properties, and regions from gallery images, detail strips, and livestream frames. SKU attribute examples specify a variant and an attribute, supply the relevant views and fields, and use the recorded value as the target; a net-weight question, for instance, must distinguish the 250 g SKU from the 500 g SKU. Product Retrieval applies the same granularity to matching. Each example pairs two listings, SKU records, or product regions with their available images and specifications, and defines equivalence at the requested level: a single 500 g bag and a two-pack of 250 g bags may depict the same product and contain the same total quantity, yet their pack configurations make them distinct SKUs. Region-level and cross-image examples extend matching to comparisons between specified visual entities.

Traceable evidence. We tie targets to the evidence that supports them so that outputs can be checked against the source listing. SKU attribute examples retain the index of the source image that contains the recorded value. Bounding-box annotations localize products, brand logos, certification marks, specification tables, and dimension annotations, and explicitly mark targets that are absent from the image. In General E-Commerce Knowledge & QA, listing images, descriptions, visible text, and structured fields are paired with questions on product knowledge, commerce OCR, and reasoning, with reference answers supported by the supplied sources. Product descriptions, entity annotations, and articles further connect product appearance with names, brands, uses, and business knowledge.

Rule-conditioned outputs. Many e-commerce decisions are defined by platform rules rather than by visual content alone. Product Information Audit pairs listing titles, gallery images, and attribute fields with platform governance rules and annotated compliance decisions. Its tasks include identifying restricted goods, detecting image–title or category inconsistencies, reviewing product-image quality, and screening content for policy violations; each target specifies the applicable rule and the decision, together with visual or textual evidence for tasks that require an explanation. The same domain also covers quality review of short product titles, where the target is a set of pass/fail judgments rather than a rewritten title. E-Commerce Content Generation applies rules to generation rather than judgment: short-title examples map structured product specifications to concise, formatted titles that follow platform style guidelines.

2.1.2 General-Purpose Data

The seven general domains preserve the perception, language, knowledge, and reasoning capabilities that e-commerce tasks build on and that the model needs beyond them. They draw on curated sources and task mixtures, including all 162 nonempty sources in OpenBee [107]. We summarize the main sources below and describe shared construction procedures in Section 2.1.3.

Perception and reading. Multimodal Understanding pairs images with Chinese or English captions and adds video data for event understanding and temporally grounded tasks, including sports event narration and general video question answering. OCR, Documents, and Charts pairs natural images, documents, charts, and tables with transcriptions, reading questions, or structured extraction targets, covering both the recognition of visible text and its use in document and chart understanding; sources include ChartNet [36], multilingual OCR collections, web articles, and table annotations. Grounding and Detection converts object and region annotations from Objects365 [73], SA-1B [35], and LocateAnything [85] into localization and detection targets conditioned on object names or referring descriptions, while GUI examples associate interface instructions with the locations of the corresponding elements.

Reasoning and knowledge. Visual Instruction & Reasoning combines single-image QA, multi-image comparison, visual quality assessment, and instruction-driven tasks. Each example pairs the available views with a question or constrained instruction and a reference answer, comparison judgment, or decision; multi-image questions require integrating evidence across views, and edit-instruction examples ask the model to infer, from a source image and its edited version, the instruction that produced the change. Risk assessment and content moderation tasks apply written platform policies to general content such as short videos; unlike Product Information Audit, they do not review product listings. Multimodal STEM pairs diagram- and video-based problems in mathematics, science, and technical reasoning with worked solutions, including material from Innovator-VL [89], and adds chart and document problems from educational corpora across biology, chemistry, mathematics, and K-12 science. Multimodal World Knowledge associates images with encyclopedic and entity-specific descriptions from online wikis and entity recaption sources, and pairs visual evidence with specialized questions whose answers draw on knowledge beyond the visible appearance.

Text-only data. Pure Text retains general records without visual inputs as language and instruction supervision. Mathematical and scientific reasoning dominate this mixture, drawing on Nemotron-Math [17] and Chinese STEM sources, while instruction-following and multilingual dialogues, Wikipedia text, reading-comprehension QA, and curated web text supply general language, factual knowledge, and instruction-following supervision.

2.1.3 Data Construction and Quality Control

Several construction and filtering procedures are shared across domains.

Spatial targets and absent objects. Conversion of box annotations in both pools makes coordinate normalization explicit, preserves the source category schemas, and keeps each target linked to its annotated visual entity. Both pools also encode absence explicitly: e-commerce annotations mark targets missing from the image, and general grounding data includes rejection examples whose queries name objects absent from the image, teaching the model to decline rather than hallucinate a box.

Recaptioning. We cluster images with WeMM-Embedding [109], sample across clusters for recaptioning, and deduplicate the resulting captions. For text-rich images, PaddleOCR [10] transcriptions provide textual cues for caption generation.

Synthesized reasoning. For synthesized STEM solutions, Qwen3.5-122B-A10B generates the reasoning and Qwen3.5-35B-A3B checks answer agreement; malformed or truncated solutions are discarded.

Web article mining. Two classifiers identify product-recommendation articles and select comparative reviews or detailed accounts of product use. Qwen3.5-122B-A10B then filters promotional or irrelevant material and extracts product and brand facts, which we deduplicate with MinHash.

Source balancing. When aggregating multi-source collections, we assign a token budget to each source group and, where a group exceeds its budget, downsample its constituent sources in proportion to their token counts, so that every source remains represented.

2.1.4 CPT Recipe

Progressive data mixture. We adopt a two-stage curriculum that increases domain specialization while maintaining broad general supervision (Figure 3). Stage 1 establishes multimodal and product representations with a 45.8% e-commerce share, and Stage 2 raises the share to 61.1%. Within the e-commerce pool (percentages are shares within each pool), supervision shifts from basic perception toward compliance auditing and knowledge-intensive QA: Product Information Audit rises from 6.4% to 18.0%, General E-Commerce Knowledge & QA from 24.7% to 31.4%, and Product Perception & Understanding falls from 37.5% to 23.6%. Within the general pool, supervision shifts from captioning and text-only data toward dense text reading

and visually grounded reasoning: OCR, Documents, and Charts rises from 13.7% to 23.8% and Multimodal STEM from 2.9% to 19.0%, while Multimodal Understanding falls from 26.4% to 8.1% and Pure Text from 26.7% to 19.2%. Throughout CPT, we track per-domain validation losses to guide mixture adjustments.

Training schedule and objective. All model sizes follow the same two-stage CPT schedule: Stage 1 trains on 370.3B tokens at a 16K sequence length, and Stage 2 extends the context window to 32K for 76.7B tokens. For a training batch $\mathcal{B}$, let v_i denote the visual inputs of example i, z_i its token sequence, and $\mathcal{M}_i$ its supervised response token positions with $N_i = |\mathcal{M}_i| > 0$. To prevent long responses from dominating gradient updates while still accounting for sequence length, we employ square averaging [8, 86], weighting each supervised token by $N_i^{-1/2}$:

$$\mathcal{L}_{\text{CPT}}(\theta; \mathcal{B}) = -\frac{\sum_{i \in \mathcal{B}} N_i^{-1/2} \sum_{t \in \mathcal{M}_i} \log \pi_\theta(z_{i,t} \mid v_i, z_{i,<t})}{\sum_{i \in \mathcal{B}} \sqrt{N_i}}. \tag{1}$$

We train with dynamic micro-batching, visual-encoder parallelism, and sequence packing up to each stage's context limit, preserving per-example loss masks and image–text alignments.

2.2 Post-Training

Following CPT, our post-training pipeline combines OPD, parameter-space model merging, and GRPO-based reinforcement learning. Both post-training stages optimize a single set of weights to support both thinking and non-thinking modes [99].

2.2.1 Post-Training Data

The post-training corpus spans general reasoning, visual perception, and instruction following alongside core e-commerce workflows such as variant matching and product detection. All model sizes share a curated pool of 499,998 OPD prompts—including specialist data for open-vocabulary detection, complex counting, and long-document QA—and 30,976 prompts for RL. We select OPD prompts via offline teacher–student comparisons, retaining instances where the teacher achieves a higher mean quality score than the student. For RL, we sample multiple rollouts from the corresponding post-OPD merged checkpoint and retain prompts with positive reward variance while filtering out already saturated instances (Appendix A.1).

2.2.2 Hybrid Thinking

To support both deliberative chain-of-thought reasoning (thinking mode) and low-latency direct responses (non-thinking mode) within a single checkpoint, both OPD and RL assign training prompts to thinking and non-thinking modes at a 4:1 ratio (Appendix A.1). Because per-response loss averaging assigns lower per-token weight to final answer tokens in longer reasoning traces, allocating a higher proportion of prompts to thinking mode balances supervision across both paradigms. During RL, all rollouts within a prompt group share the same operating mode.

2.2.3 On-Policy Distillation

OPD provides dense, token-level teacher supervision on trajectories sampled from the student policy [2]. We use a post-trained WeMM-Ecom-9B teacher for the 0.8B and 2B models, WeMM-Ecom-35B-A3B for the 4B model, and Qwen3.5-122B-A10B for the 9B and 35B-A3B models. Given a student-generated prefix $h_{i,t} = (x_i, y_{i,<t})$ of response i, the teacher evaluates the token $y_{i,t}$ sampled by the rollout policy $\pi_{\theta_{\text{old}}}$, yielding a token-level distillation signal applied identically in both modes:

$$a_{i,t}^{\text{KD}} = \log \pi_T(y_{i,t} \mid h_{i,t}) - \log \pi_{\theta_{\text{old}}}(y_{i,t} \mid h_{i,t}), \tag{2}$$

where π_T denotes the teacher policy. At any fixed prefix, the expected negative signal corresponds to the reverse KL divergence from $\pi_{\theta_{\text{old}}}$ to π_T. Holding $a_{i,t}^{\text{KD}}$ fixed during gradient updates and letting $\rho_{i,t} = \pi_\theta(y_{i,t} \mid$

$h_{i,t})/\pi_{\theta_{\text{old}}}(y_{i,t} \mid h_{i,t})$ denote the importance ratio, we minimize the clipped actor objective over batch $\mathcal{B}$ and scored token positions $\mathcal{R}_i$:

$$\mathcal{L}_{\text{OPD}}(\theta) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{|\mathcal{R}_i|}\sum_{t\in\mathcal{R}_i}\min\big(\rho_{i,t}a^{\text{KD}}_{i,t}, \text{clip}(\rho_{i,t}, 0.8, 1.2)a^{\text{KD}}_{i,t}\big). \tag{3}$$

To combine domain adaptation with the broad instruction-following priors of the backbone, we merge once after OPD and before RL by interpolating the vision and language parameters of the OPD checkpoint with the corresponding post-trained Qwen3.5 checkpoint using equal weights [11, 28].

2.2.4 Reinforcement Learning

We further refine the models via GRPO [74] using task-specific verifiable rewards. For a prompt x from task family τ, we sample G responses and compute scalar rewards $r_i = R_\tau(x, y_i)$, normalizing advantages across the group:

$$\widehat{A}_i = \frac{r_i - \bar{r}}{\text{std}(r_1, \ldots, r_G) + \epsilon_A}, \qquad \bar{r} = \frac{1}{G}\sum_i r_i. \tag{4}$$

Task rewards check predicted equivalence labels for product matching, score visual grounding and detection by weighted box IoU, and measure constraint satisfaction for instruction following (Appendix A.2). With $\ell_{i,t} = \log \pi_\theta(y_{i,t} \mid h_{i,t})$ and a frozen reference policy initialized from the pre-RL checkpoint, each online rollout batch $\mathcal{B}$ minimizes

$$\mathcal{L}_{\text{RL}}(\theta) = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{|\mathcal{R}_i|}\sum_{t\in\mathcal{R}_i}\Big[-\widehat{A}_i \exp\big(\ell_{i,t} - \text{sg}(\ell_{i,t})\big) + 0.01K_{i,t}\Big], \tag{5}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator and $K_{i,t}$ penalizes KL divergence from the reference policy (Appendix A.2).

2.3 Training Configuration

CPT employs hybrid AdamW–Muon optimization [33, 47], combining AdamW updates with Muon's orthogonalized momentum. Table 2 summarizes the training hyperparameters across all four stages for WeMM-Ecom-35B-A3B. At the 9B scale, CPT uses a peak learning rate of 2×10^{-6} and a global batch size of 2,048 sequences. Across all five model sizes, OPD and RL use a constant learning rate of 10^{-6} and a global batch size of 2,048 sampled responses.

Table 2 Training setup and hyperparameters across stages.

Setting	CPT Stage 1	CPT Stage 2	OPD	RL/GRPO
Training budget (tokens)	370.3B	76.7B	0.6B	0.2B
Optimization signal	Likelihood	Likelihood	Teacher scores	Task rewards
Optimizer	AdamW + Muon	AdamW + Muon	AdamW	AdamW
LR warmup steps	1,000	1,500	0	0
Peak learning rate	1.5×10^{-6}	10^{-6}	10^{-6}	10^{-6}
Minimum learning rate	1.5×10^{-7}	10^{-7}	10^{-6}	10^{-6}
Tokens per step	5.7M	5.7M	–	–
Global batch size	–	–	2,048	2,048
Sequence length	16,384	32,768	32,768	65,536

3 E-Commerce Evaluation Framework

Evaluating multimodal models for real-world commerce requires testing whether they can interpret complex product listings, align variants across views, enforce platform governance rules, and localize advertised items in livestream frames. We construct a comprehensive evaluation suite of 28 benchmarks—comprising 23 in-house benchmarks and five public benchmarks, totaling 207,117 evaluation instances—organized into the eight core

scenarios illustrated in Figure 4. Because the in-house benchmarks derive from proprietary business data, we detail their task definitions, input modalities, and scoring metrics in Appendix B.

To prevent data contamination between training and business evaluation splits, we enforce strict deduplication across splits using product identifiers, cryptographic hashes for exact image duplicates, perceptual hashes for near duplicates, and visual embedding similarity.

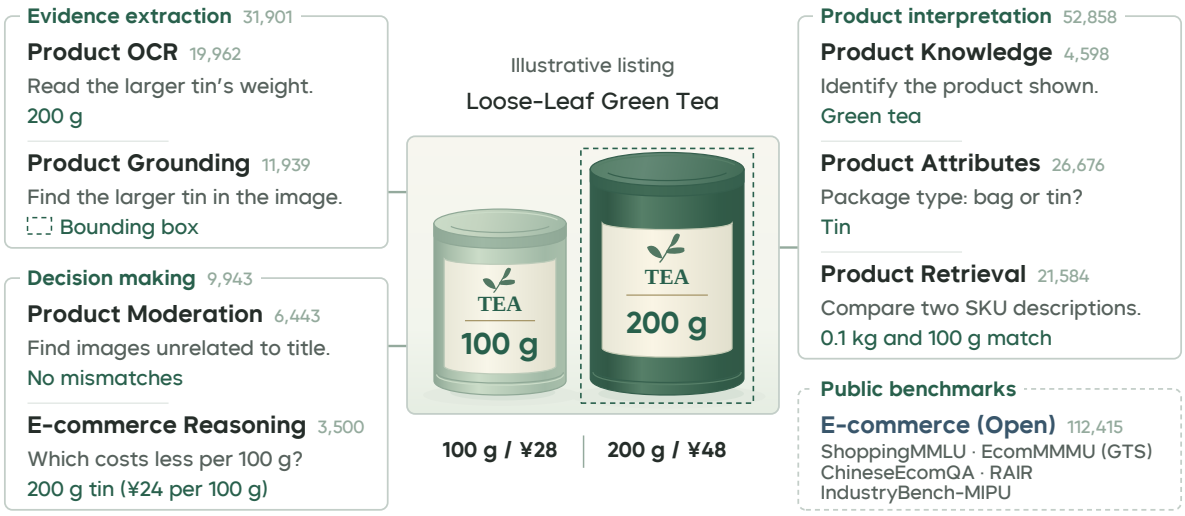


Figure 4 Overview of the eight e-commerce evaluation scenarios. Seven in-house scenarios are organized around an illustrative product listing into evidence extraction, product interpretation, and decision making, with representative prompts and target outputs (green), complemented by five public benchmarks in E-commerce (Open). Numbers in gray denote instance counts.

Evidence extraction (Product OCR and Product Grounding). Product OCR (19,962 instances) evaluates fine-grained text recognition and specification reading across single product photos, multi-angle gallery views, and long promotional strips. Product Grounding (six benchmarks, 11,939 instances) assesses spatial localization of target products (Product Detection), brand logos (Logo Detection), certification marks (Qualification Detection), specification tables (Chart Detection), dimension annotations (Dimension Detection), and advertised items in livestream frames (Livestream Detection, 1,992 instances).

Product interpretation (Product Knowledge, Product Attributes, and Product Retrieval). Product Knowledge (two benchmarks, 4,598 instances) tests fine-grained entity recognition and domain knowledge across products, brands, and industry standards. Product Attributes (four benchmarks, 26,676 instances) evaluates attribute value extraction, brand attribution, and multi-choice property selection from interleaved images and structured fields. Product Retrieval (three benchmarks, 21,584 instances) benchmarks cross-item matching at three granularity levels: judging visual similarity between cropped regions (Region-level), determining whether two multimodal listings represent the same product (Product-level), and verifying exact variant equivalence from SKU specifications (SKU-level).

Decision making (Product Moderation and E-commerce Reasoning). Product Moderation (three benchmarks, 6,443 instances) evaluates rule-grounded compliance decisions, including detecting prohibited goods, enforcing category sales restrictions, and identifying image–title mismatches. E-commerce Reasoning (four benchmarks, 3,500 instances) comprises single-image analytical tasks covering chart reasoning, taxonomy classification, constrained product counting, and spatial relation reasoning, where reference answers are programmatically derived from the underlying product tables and query constraints.

Public benchmark coverage. The E-commerce (Open) scenario incorporates 112,415 instances from five public benchmarks: ShoppingMMLU and ChineseEcomQA for e-commerce concepts and operations [4, 32], RAIR for rule-aware search relevance [48], and EcomMMMU and IndustryBench-MIPU for multimodal product understanding and multi-image industrial attribute extraction [41, 63].

4 Results

4.1 Experimental Setup

We evaluate all five WeMM-Ecom models in both thinking and non-thinking modes, using a single shared checkpoint per model scale across both operating modes and across both evaluation suites. Our primary comparison pairs each WeMM-Ecom model with the corresponding post-trained Qwen3.5 baseline of the same parameter scale and mode across all 28 e-commerce benchmarks and 64 general-purpose benchmarks.

Evaluation protocol. We report the unweighted mean of primary metrics across the 28 e-commerce benchmarks and, separately, across the 64 general benchmarks on a 0–100 scale (where RefCOCO, RefCOCO+, and RefCOCOg are averaged into a single RefCOCO score and the English and Chinese splits of MMBench count as separate benchmarks). Tasks requiring open-ended semantic adjudication use Qwen3-30B-A3B-Instruct-2507 [95] as an independent judge, distinct from the Qwen3.5-122B-A10B verifier used during RL training. In flagship comparisons, each external baseline is reported under its best-performing evaluated operating mode, i.e., whichever of thinking and non-thinking mode scores higher for that model. Detailed scoring protocols and decoding settings are provided in Appendix B.

4.2 E-Commerce Results

Flagship comparison. Table 3 compares WeMM-Ecom-35B-A3B against leading open-source flagships, including Qwen3.8-Flash-Next [66], HY4 [80], MiMo-V2.6-Flash-RL [65], DeepSeek-V4.1-Flash-1017 [12], and GLM-5.3-Flash [103]. Despite having 5×–22× fewer total parameters, WeMM-Ecom-35B-A3B achieves the highest overall average of 74.4 in non-thinking mode, outperforming the strongest baseline, Qwen3.8-Flash-Next (73.2), by about 1.2 points. Substantial gains in Product Retrieval and Product Grounding drive this overall lead (Table 3), which remains consistent under alternative scenario weighting schemes.

Dense models. Table 4 reports thinking and non-thinking results for all four dense WeMM-Ecom models alongside Qwen3.5-9B and the e-commerce specialist Valley3-32B [7]. WeMM-Ecom-9B achieves overall means of 72.5/72.1 (thinking/non-thinking), surpassing Qwen3.5-9B (67.3/65.7) by 5.2/6.4 points, Valley3-32B (64.6/68.6) by 7.9/3.5 points, and the 3× larger Qwen3.8-27B [68] (71.8; Figure 1) by 0.7 points. At the compact scales, WeMM-Ecom-4B, 2B, and 0.8B reach 68.7/68.7, 64.2/63.3, and 57.6/56.7, respectively. Across all five model sizes and both operating modes, WeMM-Ecom consistently outperforms size-matched Qwen3.5, with the largest non-thinking improvements reaching +15.5 and +12.3 points at 0.8B and 2B.

4.3 Analysis of E-Commerce Gains

Fine-grained visual localization and cross-listing variant matching account for the largest capability leaps on the e-commerce suite. At the 9B scale, WeMM-Ecom improves upon Qwen3.5-9B across all eight scenario means in thinking mode and seven of eight in non-thinking mode. Product Grounding improves by 15.7 points in thinking mode and 10.3 points in non-thinking mode, while Product Retrieval gains 13.4 points in non-thinking mode—directly reflecting the bounding-box localization and multi-level equivalence supervision introduced in Section 2.1.1.

Broad improvements beyond grounding. To confirm that WeMM-Ecom's advantage extends well beyond spatial localization, we re-evaluate the suite under alternative subsets and weighting schemes. Excluding all six Product Grounding benchmarks retains clear positive gains over size-matched Qwen3.5 across all ten size–mode pairs—including non-thinking improvements of +11.1 and +8.7 points at 0.8B and 2B—and leaves WeMM-Ecom-35B-A3B with the highest mean among the models in Table 3 (75.7 for WeMM-Ecom-35B-A3B

Table 3 Comparison of WeMM-Ecom-35B-A3B with flagship open-source baselines on e-commerce benchmarks. Bold and underlining mark the highest and second-highest scores in each row.

Benchmark	WeMM-Ecom	Qwen3.8 Flash-Next	HY4	DeepSeek V4.1-Flash-1017	GLM-5.3 Flash	MiMo-V2.6 Flash-RL
Size	**35B-A3B**	**180B-A6B**	**770B-A49B**	**763B-A16B**	**320B-A18B**	**310B-A15B**
Product Knowledge						
E-commerce Knowledge	67.1	78.9	**85.2**	82.4	82.9	79.2
Entity Recognition	80.4	84.2	**86.7**	81.6	84.9	76.2
Product Attributes						
Attribute Selection	**77.1**	75.7	76.5	74.0	71.9	68.3
Attribute Extraction	**81.8**	78.8	77.9	79.0	76.4	76.9
Brand Attribution	83.1	**86.3**	84.3	82.5	80.5	80.8
Brand Extraction	**83.6**	81.2	80.5	82.7	83.1	81.5
Product Moderation						
Illegal Goods	93.0	**93.7**	89.0	**93.7**	93.1	93.0
Sales Restrictions	**91.1**	89.7	90.6	86.6	88.4	90.2
Image-Title Mismatch	87.4	**91.7**	89.7	87.7	89.5	84.0
Product OCR	**97.8**	97.7	97.1	95.9	97.0	93.8
Product Retrieval						
Region-level Retrieval	**85.1**	78.6	76.0	83.9	76.7	76.1
Product-level Retrieval	**87.5**	66.5	63.9	78.3	61.9	66.1
SKU-level Retrieval	**62.2**	43.6	43.0	49.3	49.3	38.7
Product Grounding						
Chart Detection	77.0	**79.4**	72.7	66.5	65.4	27.6
Product Detection	**79.9**	71.7	73.3	67.1	67.6	44.3
Logo Detection	58.9	**59.3**	49.2	42.2	36.0	6.1
Qualification Detection	**61.8**	54.4	37.2	23.6	18.1	2.0
Dimension Detection	**74.3**	**74.3**	68.1	55.7	55.4	26.8
Livestream Detection	**64.7**	47.7	55.6	53.0	50.1	27.1
E-commerce Reasoning						
Chart Reasoning	99.4	**99.5**	99.1	93.5	98.5	96.8
Category Reasoning	81.2	85.6	**85.8**	85.2	85.6	80.2
Product Counting	82.2	79.6	81.2	69.0	**83.8**	72.8
Spatial Reasoning	31.2	**35.6**	32.6	19.4	35.4	17.4
E-commerce (Open)						
ShoppingMMLU [32]	74.9	78.0	76.9	76.0	**78.3**	76.3
RAIR [48]	43.7	**48.0**	45.0	44.9	45.9	45.9
ChineseEcomQA [4]	64.2	71.6	**74.7**	68.2	70.5	69.3
EcomMMMU (GTS) [41]	48.0	50.3	**54.9**	49.3	53.3	48.9
IndustryBench-MIPU [63]	63.5	66.8	**72.8**	63.2	69.7	65.3
Overall mean (28)	**74.4**	73.2	72.1	69.1	69.6	61.1

Each comparison model is reported in its best-performing evaluated operating mode (thinking or non-thinking).

versus 75.5 for Qwen3.8-Flash-Next, 75.6 for HY4, and 73.9 for DeepSeek-V4.1-Flash-1017, keeping each model in its Table 3 operating mode). Likewise, WeMM-Ecom maintains positive gains across all ten size–mode comparisons when weighting all eight scenarios equally (+2.1 to +14.8 points, which also preserves the flagship ranking), when omitting any single scenario (at least +0.9 points), or when excluding the two probability-readout retrieval sets (+3.1 to +17.8 points).

Scenario-level trade-offs and format robustness. While overall e-commerce performance improves across scales, gains vary by task family: Qwen3.5-35B-A3B is slightly higher on Product Moderation in thinking mode, and Qwen3.5-9B slightly leads WeMM-Ecom-9B on E-commerce Reasoning in non-thinking mode. For structured localization tasks such as Chart Detection, our evaluation parser accepts class predictions

Table 4 Comparison of dense WeMM-Ecom models with baselines on e-commerce benchmarks. Results are thinking / non-thinking. Bold and underlining mark the highest and second-highest scores in each row. Region-level and Product-level Retrieval are scored differently in thinking and non-thinking modes; see Appendix B.

Benchmark	**WeMM-Ecom**				**Qwen3.5**	**Valley3**
Size	**0.8B**	**2B**	**4B**	**9B**	**9B**	**32B**
			Product Knowledge			
E-commerce Knowledge	26.7 / 30.8	36.4 / 38.6	50.9 / 51.4	58.7 / 59.4	57.5 / 55.8	**63.4** / 62.8
Entity Recognition	53.1 / 66.9	74.4 / 76.6	78.4 / 78.4	80.2 / 77.3	79.8 / 75.6	79.3 / **80.6**
			Product Attributes			
Attribute Selection	66.2 / 62.6	71.1 / 71.3	73.3 / 74.1	73.6 / **74.5**	71.0 / 70.1	69.2 / 71.9
Attribute Extraction	71.8 / 75.2	77.6 / 79.2	79.9 / 80.4	80.8 / **81.1**	77.0 / 73.6	74.1 / 78.6
Brand Attribution	79.1 / 79.2	82.6 / 83.3	81.8 / 80.8	**83.5** / 81.0	79.8 / 73.6	77.0 / 80.3
Brand Extraction	80.1 / 80.8	81.0 / 82.5	80.8 / 83.2	81.2 / **85.1**	81.8 / 81.8	83.1 / 83.8
			Product Moderation			
Illegal Goods	83.8 / 84.2	88.4 / 87.4	91.5 / 91.5	**91.9** / 91.7	91.0 / 88.4	91.8 / 90.3
Sales Restrictions	82.6 / 77.5	84.3 / 83.1	89.3 / 90.0	89.9 / **90.8**	89.7 / 83.3	90.4 / 88.2
Image-Title Mismatch	77.2 / 38.5	82.7 / 62.3	88.3 / 82.2	85.7 / 85.6	80.2 / 70.1	**89.1** / 86.7
Product OCR	93.3 / 93.8	96.3 / 96.2	97.1 / 97.5	**97.6** / 97.3	97.2 / 96.4	95.1 / 96.1
			Product Retrieval			
Region-level Retrieval	74.2 / 83.5	78.8 / 84.4	80.2 / 84.3	80.7 / **85.6**	76.9 / 80.7	73.4 / 81.6
Product-level Retrieval	64.3 / 85.3	60.9 / **88.5**	64.7 / 85.3	61.1 / 85.8	61.6 / 79.9	67.9 / 83.6
SKU-level Retrieval	53.7 / 63.1	62.4 / **64.8**	58.3 / 64.0	61.9 / 61.4	52.1 / 32.0	33.0 / 43.2
			Product Grounding			
Chart Detection	52.4 / 50.6	64.4 / 55.7	61.4 / 66.9	**73.4** / 73.3	60.8 / 63.6	69.4 / 65.4
Product Detection	74.8 / 75.1	68.6 / 72.3	76.4 / 78.5	79.2 / **80.4**	70.0 / 72.3	66.0 / 71.8
Logo Detection	45.5 / 41.8	44.2 / 43.1	54.1 / 51.1	**59.1** / 54.0	44.5 / 43.4	41.5 / 42.0
Qualification Detection	16.6 / 19.7	24.7 / 25.1	27.9 / 35.5	49.0 / **51.4**	22.6 / 37.2	34.3 / 40.2
Dimension Detection	30.5 / 42.2	44.4 / 38.7	40.5 / 36.8	77.2 / **85.9**	55.9 / 75.2	41.1 / 73.2
Livestream Detection	55.9 / 58.5	59.9 / 60.9	59.2 / **61.8**	58.6 / 60.2	48.5 / 52.0	40.9 / 47.1
			E-commerce Reasoning			
Chart Reasoning	77.6 / 65.9	96.3 / 94.4	99.2 / 98.3	**99.4** / 95.9	98.7 / 98.4	73.6 / 89.5
Category Reasoning	77.6 / 76.6	80.2 / 80.4	83.4 / 82.2	**84.2** / 82.4	83.0 / 80.6	81.6 / 81.6
Product Counting	38.0 / 34.4	63.2 / 56.0	79.8 / 70.8	**80.4** / 68.6	77.4 / 70.2	60.6 / 74.2
Spatial Reasoning	3.8 / 1.2	13.4 / 10.6	32.4 / 24.4	**38.2** / 21.6	29.2 / 20.6	16.4 / 17.4
			E-commerce (Open)			
ShoppingMMLU [32]	61.1 / 50.2	70.5 / 60.6	74.9 / 69.7	**76.6** / 72.5	76.0 / 69.4	73.3 / 73.8
RAIR [48]	37.0 / 23.8	36.1 / 36.6	40.5 / 36.1	44.3 / 42.7	44.6 / 41.5	**48.7** / 44.5
ChineseEcomQA [4]	50.6 / 45.4	57.6 / 46.9	61.1 / 61.1	64.3 / 60.0	61.4 / 54.8	59.8 / **66.9**
EcomMMMU (GTS) [41]	40.5 / 41.5	41.1 / 41.7	50.6 / 50.4	51.5 / **52.0**	48.6 / 42.0	46.8 / 48.5
IndustryBench-MIPU [63]	44.9 / 39.7	57.6 / 51.9	**67.1** / 57.0	66.5 / 61.4	66.6 / 56.1	**67.1** / 57.4
Overall mean (28)	57.6 / 56.7	64.2 / 63.3	68.7 / 68.7	**72.5** / 72.1	67.3 / 65.7	64.6 / 68.6

under both `category` and `label` keys so that comparisons measure localization quality rather than JSON key conventions (Appendix B). Even under this schema-compatible parser, WeMM-Ecom-35B-A3B outperforms Qwen3.5-35B-A3B on Chart Detection by 8.1 points in thinking mode and 13.0 points in non-thinking mode.

4.4 General Benchmarks

To assess general-purpose perception and reasoning, we evaluate models across 64 public benchmarks comprising 54 multimodal benchmarks—such as MMBench, MMMU, and MathVista [45, 49, 101]—and 10 text-only benchmarks, including MMLU-Pro and IFEval [87, 108].

Comparison with state-of-the-art models. Table 5 compares WeMM-Ecom-35B-A3B with Qwen3.5-35B-A3B [69], Qwen3.6-35B-A3B [67], MiniCPM-V-4.7-35B-A3B [99], and Gemma4-26B-A4B-it [78]. In thinking mode, WeMM-Ecom-35B-A3B achieves the highest overall average of 76.4, edging out both Qwen3.5-35B-A3B (76.2) and Qwen3.6-35B-A3B (76.3). In the corresponding non-thinking comparison, WeMM-Ecom-35B-A3B scores 72.6, outperforming Qwen3.5-35B-A3B (71.0) by 1.6 points.

Table 5 Comparison of WeMM-Ecom-35B-A3B with comparable baselines on general benchmarks. Results are thinking / non-thinking. Bold and underlining mark the highest and second-highest scores in each row.

Benchmark	WeMM-Ecom	Qwen3.5	Qwen3.6	MiniCPM-V-4.7	Gemma4
Size	**35B-A3B**	**35B-A3B**	**35B-A3B**	**35B-A3B**	**26B-A4B**
STEM and Multimodal Reasoning					
CMMMU (val) [105]	73.7 / 69.8	73.6 / 69.1	**73.8** / 67.1	70.6 / 70.0	59.3 / 60.9
CMMU (val, MCQ) [23]	93.2 / 91.9	**93.4** / 91.4	92.3 / 91.1	92.7 / 92.8	71.7 / 74.8
CSVQA [30]	83.5 / 73.6	83.3 / 75.3	**84.2** / 74.2	77.3 / 81.8	43.5 / 48.5
DynaMath [111]	**84.6** / 83.5	81.8 / 79.8	81.5 / 77.6	81.6 / 81.7	64.6 / 72.9
LogicVista [93]	78.0 / 74.0	**79.0** / 72.0	78.0 / 71.0	69.0 / 72.0	63.0 / 67.0
MATH-Vision [84]	79.6 / 72.4	80.9 / 72.6	**81.7** / 73.3	66.0 / 72.2	64.2 / 65.4
MathVista (mini) [49]	**87.5** / 85.7	85.8 / 83.7	86.2 / 84.3	85.5 / 86.5	75.9 / 76.6
MMMU [101]	80.1 / 79.3	80.8 / 78.6	**82.9** / 79.1	78.3 / 79.1	74.4 / 73.1
MMMU-Pro [102]	74.1 / 69.2	**74.5** / 69.5	**74.5** / 69.0	62.7 / 66.1	47.3 / 56.1
VisualPuzzles [75]	**61.8** / 54.8	60.3 / 53.5	61.6 / 53.3	52.7 / 56.3	46.0 / 53.0
VisuLogic [94]	31.7 / 32.6	32.3 / 31.2	**33.3** / 28.1	29.5 / 32.7	11.2 / 27.7
ZeroBench (sub) [71]	35.0 / 31.7	34.7 / 30.2	**36.8** / 29.9	25.5 / 26.7	11.7 / 21.0
General VQA					
CCBench [59]	78.2 / 79.8	76.1 / 77.5	76.5 / 73.9	**82.4** / 81.2	61.4 / 61.6
ChineseSimpleVQA [22]	**58.8** / 56.1	56.4 / 54.2	55.2 / 54.7	52.1 / 52.0	53.1 / 47.4
CII-Bench [104]	**70.9** / 67.0	67.6 / 67.1	69.8 / 68.6	66.6 / 68.8	67.1 / 69.6
$\text{MMBench}_{\text{EN-DEV-v1.1}}$ [45]	90.5 / 89.3	90.6 / 88.5	90.6 / 88.4	91.3 / **92.0**	86.1 / 86.7
$\text{MMBench}_{\text{CN-DEV-v1.1}}$ [45]	90.1 / 89.5	90.1 / 88.6	89.6 / 87.2	**91.3** / 88.9	85.6 / 84.9
MMStar [6]	**79.1** / 78.1	79.0 / 77.7	78.9 / 78.0	77.1 / 77.5	73.0 / 72.2
MM-Vet [100]	**89.4** / 86.0	85.9 / 82.5	87.1 / 82.1	83.4 / 84.3	79.1 / 76.6
PerceptionBench [40]	41.5 / 38.4	**43.6** / 37.3	42.9 / 36.6	34.2 / 35.8	32.4 / 34.4
RealWorldQA [92]	83.0 / 78.7	**84.7** / 77.3	83.1 / 81.7	76.3 / 76.6	65.8 / 67.2
SimpleVQA [9]	**67.6** / 60.8	62.9 / 59.5	60.7 / 58.3	61.0 / 58.4	56.4 / 54.0
WorldVQA [110]	23.4 / 24.0	26.8 / 23.1	**27.2** / 24.8	15.7 / 16.1	14.9 / 14.0
Documents and Charts					
AI2D (test) [34]	89.4 / 89.3	90.4 / 88.9	**91.0** / 89.4	90.3 / 89.6	83.9 / 84.7
CC-OCR [97]	85.1 / **86.3**	85.5 / 86.2	85.2 / 86.1	77.3 / 81.6	65.0 / 76.0
ChartQAPro [55]	**68.9** / 67.2	68.2 / 66.3	68.2 / 64.4	64.4 / 65.6	58.9 / 63.0
ChartQA (test) [54]	91.0 / 88.4	**91.2** / 86.8	91.0 / 87.0	89.4 / 86.9	87.5 / 83.4
CharXiv (RQ) [88]	76.8 / 70.4	**77.8** / 69.9	75.0 / 69.6	67.9 / 68.3	63.6 / 62.8
DocVQA (val) [56]	**97.7** / 96.9	97.5 / 96.6	97.3 / 96.0	96.5 / 94.6	89.9 / 90.3
InfographicVQA (val) [57]	92.8 / 91.2	93.0 / 90.7	**93.2** / 91.3	85.8 / 84.9	70.7 / 67.0
OCRBench [46]	**90.9** / 90.6	88.7 / 89.5	88.5 / 90.1	87.5 / 89.1	82.9 / 83.6
OCRBench v2 [19]	64.6 / 63.9	65.5 / 64.1	**66.8** / 66.5	49.5 / 61.8	39.1 / 44.3
OmniDocBench [60]	90.6 / **92.7**	89.6 / 88.6	91.5 / 90.3	77.6 / 92.0	3.1 / 27.0
WildDoc [82]	68.6 / 70.9	**77.0** / 74.8	76.8 / 74.4	67.3 / 70.5	43.3 / 48.7
Grounding and Counting					
CountBenchQA [3]	96.5 / 91.0	**96.7** / 92.1	96.1 / 92.1	92.9 / 90.4	95.1 / 85.7
CountQA [76]	**70.2** / 43.9	69.4 / 40.1	67.3 / 40.1	35.5 / 31.7	39.3 / 26.6
MIHBench [37]	**83.1** / 79.1	82.4 / 60.0	81.6 / 63.9	81.3 / 81.5	75.8 / 44.0
ODInW13 [38]	48.6 / 44.5	45.0 / 48.8	**50.8** / 50.4	26.3 / 27.8	35.3 / 33.0
PixMo-Count [13]	89.9 / 75.4	**90.1** / 75.8	89.7 / 73.5	71.0 / 69.3	81.9 / 62.4
RefCOCO (avg) [53, 98]	88.6 / 88.9	88.4 / 89.2	88.1 / **89.3**	88.7 / 88.9	79.1 / 79.0
Ref-L4 (val) [5]	90.0 / **90.5**	89.2 / 88.1	88.8 / 88.3	77.8 / 86.5	76.4 / 86.0
TallyQA [1]	**83.1** / 82.6	80.3 / 81.5	79.7 / 81.3	81.2 / 80.6	78.8 / 79.0

Continued on next page

Table 5 (continued)

Benchmark	WeMM-Ecom	Qwen3.5	Qwen3.6	MiniCPM-V-4.7	Gemma4
Size	**35B-A3B**	**35B-A3B**	**35B-A3B**	**35B-A3B**	**26B-A4B**
2D/3D Spatial Understanding					
CV-Bench [81]	**88.9** / 88.4	88.0 / 87.8	87.7 / 87.9	86.8 / 86.5	84.2 / 80.0
EmbSpatial-Bench [16]	83.7 / 78.9	83.9 / 79.3	**84.1** / 79.3	82.7 / 82.2	78.5 / 70.3
ERQA [77]	56.5 / 48.8	**61.0** / 50.3	58.5 / 52.3	55.0 / 49.8	47.0 / 45.3
MMSI-Bench [96]	39.1 / 32.8	40.2 / 32.2	**41.4** / 32.4	29.4 / 35.1	27.1 / 28.8
OmniSpatial [29]	**61.0** / 55.5	60.6 / 53.0	59.8 / 54.1	53.5 / 53.6	49.3 / 44.9
SpatialMQA [44]	58.5 / 47.5	**58.9** / 45.3	58.1 / 47.1	50.9 / 47.5	57.8 / 44.5
Multi-Image Understanding					
BLINK [20]	69.7 / 68.6	69.7 / 66.1	**70.4** / 65.6	66.4 / 67.0	58.3 / 63.0
Mantis-Eval [31]	88.5 / 86.6	87.6 / 84.8	**90.3** / 87.6	87.6 / 86.6	83.0 / 74.7
MMIU [58]	66.2 / 66.2	67.6 / 63.8	**70.1** / 64.1	64.4 / 63.5	53.4 / 50.1
MuirBench [83]	70.9 / 71.0	73.2 / 62.0	**74.7** / 73.7	72.2 / 65.5	57.8 / 60.0
Instruction Following					
CFBench [106]	86.2 / 84.4	85.2 / 84.9	**87.0** / 84.1	86.2 / 85.6	78.6 / 82.8
IFBench [62]	**75.7** / 51.0	71.3 / 43.3	65.3 / 35.7	55.3 / 42.3	64.7 / 43.0
IFEval [108]	**94.3** / 86.7	93.7 / 85.4	91.7 / 83.2	91.1 / 88.4	87.3 / 87.8
MIA-Bench [64]	**87.2** / 83.8	86.8 / 82.8	**87.2** / 85.1	83.7 / 86.2	83.1 / 81.3
MM-IFEval [15]	82.5 / 74.3	**82.8** / 65.2	80.9 / 66.3	69.9 / 70.5	69.8 / 69.6
Text-Centric Tasks					
C-Eval [26]	**91.4** / 89.7	91.1 / 90.4	90.8 / 90.7	90.1 / 89.9	79.4 / 74.9
GPQA (Diamond) [70]	81.8 / 77.8	81.3 / **86.4**	81.8 / 84.9	74.8 / 76.8	64.7 / 63.6
HMMT (Nov. 2025) [14]	87.5 / 84.2	86.7 / 69.2	**88.3** / 67.5	78.3 / 65.8	66.7 / 64.2
IMO-AnswerBench (v2) [51]	73.3 / 62.5	74.0 / 58.3	**75.3** / 59.8	67.3 / 56.5	52.5 / 50.8
MMLU-Pro [87]	85.3 / 84.1	**85.7** / 84.3	85.3 / 84.5	84.5 / 84.6	65.7 / 80.9
MMLU-Redux [21]	**93.5** / 91.3	**93.5** / 88.8	93.4 / **93.5**	93.3 / 93.4	86.0 / 90.2
SuperGPQA [18]	64.3 / 61.1	64.8 / 62.5	**65.9** / 63.7	60.5 / 61.3	44.0 / 52.1
Overall mean (64)	**76.4** / 72.6	76.2 / 71.0	76.3 / 71.2	70.5 / 70.8	62.4 / 62.4

Dense models. Table 6 presents thinking and non-thinking results for our four dense models alongside Qwen3.5-9B and GLM-4.6V-Flash [24]. WeMM-Ecom-9B averages 74.2/69.7 (thinking/non-thinking), outperforming Qwen3.5-9B (73.4/67.6) by 0.8/2.1 points and GLM-4.6V-Flash (63.3/57.9) by 10.8/11.8 points. At the compact scales (4B, 2B, and 0.8B), WeMM-Ecom surpasses size-matched Qwen3.5 in thinking mode by 0.4, 4.5, and 6.9 points, respectively. As shown in Figure 1, WeMM-Ecom-2B ranks first among lightweight models in non-thinking mode.

Robustness against published baselines. As a stringent external check, we also compare WeMM-Ecom against an upper-bound Qwen3.5 baseline constructed by taking the higher of our local evaluation and the official Qwen3.5 model-card score [69] on every comparable benchmark. Even against these composite upper bounds (46.2/44.2 at 0.8B and 59.9/55.7 at 2B in thinking/non-thinking modes), WeMM-Ecom-0.8B and WeMM-Ecom-2B maintain substantial overall leads of +6.1/+3.8 and +3.6/+4.7 points, respectively, demonstrating that domain-grounded training and distillation significantly enhance general capabilities in compact models.

Table 6 Comparison of dense WeMM-Ecom models with baselines on general benchmarks. Results are thinking / non-thinking. Bold and underlining mark the highest and second-highest scores in each row.

Benchmark	WeMM-Ecom				Qwen3.5	GLM-4.6V-Flash
Size	**0.8B**	**2B**	**4B**	**9B**	**9B**	**9B**
STEM and Multimodal Reasoning						
CMMMU (val) [105]	41.8 / 34.4	55.6 / 51.3	65.4 / 59.9	69.9 / 66.2	**70.9** / 65.1	61.1 / 53.0
CMMU (val, MCQ) [23]	53.7 / 46.3	76.2 / 72.4	89.1 / 84.3	91.2 / 88.6	**91.8** / 87.9	80.4 / 70.7
CSVQA [30]	25.2 / 25.8	51.5 / 45.9	71.5 / 61.4	**79.9** / 69.7	77.9 / 68.8	52.3 / 36.6
DynaMath [111]	60.0 / 49.2	76.8 / 71.8	82.4 / 78.3	**83.9** / 80.3	81.4 / 74.0	73.6 / 63.5
LogicVista [93]	48.0 / 39.0	59.0 / 57.0	70.0 / 67.0	**77.0** / 71.0	75.0 / 68.0	60.0 / 52.0
MATH-Vision [84]	27.4 / 27.5	52.2 / 50.0	70.1 / 65.2	76.3 / 70.9	**77.7** / 69.7	55.9 / 42.1
MathVista (mini) [49]	68.1 / 66.1	80.2 / 79.0	86.2 / 82.6	**86.5** / 83.9	84.6 / 83.3	83.0 / 72.7
MMMU [101]	54.0 / 42.7	65.7 / 63.3	75.6 / 72.7	**79.1** / 76.7	78.7 / 75.1	72.3 / 67.2
MMMU-Pro [102]	35.1 / 31.5	51.9 / 49.8	65.4 / 62.4	**69.8** / 64.9	69.3 / 66.0	56.9 / 47.8
VisualPuzzles [75]	32.3 / 31.3	40.4 / 41.7	49.2 / 50.2	56.9 / 53.8	**58.1** / 49.0	42.0 / 42.6
VisuLogic [94]	24.0 / 25.7	26.6 / 25.8	29.4 / 26.7	32.0 / 28.9	**34.2** / 29.7	26.6 / 15.3
ZeroBench (sub) [71]	15.3 / 11.4	21.9 / 18.3	26.7 / 29.0	**33.5** / 28.1	32.3 / 29.3	24.0 / 16.5
General VQA						
CCBench [59]	40.0 / 45.3	60.0 / 64.7	66.1 / 69.4	**74.7** / 74.5	69.6 / 70.2	67.8 / 61.4
ChineseSimpleVQA [22]	17.6 / 18.9	32.1 / 29.4	45.5 / 41.4	49.7 / 48.5	44.3 / 45.7	**50.2** / 44.6
CII-Bench [104]	38.4 / 37.8	48.9 / 48.4	61.3 / 56.6	**65.6** / 64.4	**65.6** / 63.6	57.0 / 54.8
$\text{MMBench}_{\text{EN-DEV-v1.1}}$ [45]	73.1 / 73.0	85.0 / 83.0	87.5 / 86.7	**90.2** / 89.6	88.2 / 86.2	86.2 / 82.8
$\text{MMBench}_{\text{CN-DEV-v1.1}}$ [45]	72.0 / 70.3	83.5 / 81.6	87.6 / 87.3	**89.4** / 88.9	87.2 / 85.4	86.3 / 83.0
MMStar [6]	62.9 / 58.0	72.3 / 71.3	76.5 / 75.7	**80.2** / 77.6	77.7 / 74.6	73.2 / 68.1
MM-Vet [100]	66.0 / 62.6	75.6 / 72.0	81.5 / 76.3	**85.9** / 80.3	80.0 / 76.9	77.4 / 76.1
PerceptionBench [40]	25.2 / 23.6	31.3 / 29.0	36.5 / 32.8	**41.4** / 36.5	39.6 / 33.9	31.2 / 28.3
RealWorldQA [92]	68.1 / 66.1	75.6 / 71.6	77.7 / 78.2	**81.1** / 77.0	80.5 / 74.1	75.8 / 72.2
SimpleVQA [9]	42.7 / 37.9	54.8 / 48.3	59.1 / 53.7	**59.9** / 57.6	56.4 / 53.8	58.4 / 53.3
WorldVQA [110]	8.2 / 6.5	14.1 / 13.8	20.6 / 16.7	**22.1** / 20.1	20.7 / 19.4	14.9 / 12.1
Documents and Charts						
AI2D (test) [34]	73.9 / 73.1	83.7 / 82.7	87.8 / 86.2	**89.4** / 88.3	**89.4** / 86.8	84.7 / 84.8
CC-OCR [97]	73.6 / 77.6	80.6 / 81.8	82.4 / 84.6	84.7 / **85.6**	84.3 / 85.3	63.7 / 62.8
ChartQAPro [55]	49.0 / 33.8	58.4 / 56.2	65.6 / 64.7	**68.2** / 66.8	67.0 / 64.3	58.1 / 43.4
ChartQA (test) [54]	87.0 / 82.6	89.7 / 87.1	90.6 / 87.2	**90.8** / 87.8	89.9 / 86.6	89.0 / 88.2
CharXiv (RQ) [88]	50.7 / 45.3	60.9 / 59.5	70.3 / 64.4	73.1 / 68.6	**74.3** / 67.9	60.7 / 50.3
DocVQA (val) [56]	94.1 / 93.8	96.2 / 96.4	97.1 / 96.6	**97.5** / 97.0	97.3 / 96.1	96.5 / 96.5
InfographicVQA (val) [57]	77.8 / 72.8	87.7 / 84.6	90.8 / 87.9	**92.2** / 90.9	91.8 / 88.3	89.0 / 86.2
OCRBench [46]	80.6 / 84.6	88.0 / 89.0	86.2 / 86.7	89.8 / **90.6**	87.8 / 85.2	85.4 / 85.3
OCRBench v2 [19]	50.1 / 48.2	56.3 / 51.3	60.8 / 57.9	**65.6** / 60.9	64.0 / 62.1	53.2 / 50.3
OmniDocBench [60]	80.1 / 84.3	85.9 / 86.9	86.1 / 90.7	**90.9** / 90.5	86.8 / 88.6	53.5 / 50.6
WildDoc [82]	61.3 / 54.7	67.7 / 52.8	67.7 / 65.5	70.4 / 69.2	**71.8** / 70.4	70.0 / 64.6
Grounding and Counting						
CountBenchQA [3]	75.8 / 72.1	90.0 / 85.1	93.3 / 84.5	95.5 / 90.8	**97.0** / 87.8	90.6 / 83.1
CountQA [76]	18.3 / 12.4	37.0 / 31.2	46.8 / 33.3	**65.3** / 38.6	61.8 / 37.4	35.9 / 27.6
MIHBench [37]	78.6 / 42.6	83.1 / 53.5	**83.6** / 52.4	82.6 / 56.9	82.6 / 51.9	81.1 / 80.6
ODInW13 [38]	33.0 / 26.4	40.5 / 37.9	41.8 / 40.6	**44.5** / 43.2	41.8 / 42.8	26.3 / 26.2
PixMo-Count [13]	52.9 / 42.0	75.6 / 70.0	80.9 / 63.7	**87.8** / 68.3	87.6 / 74.2	73.9 / 57.6
RefCOCO (avg) [53, 98]	76.8 / 78.8	82.8 / 81.9	85.2 / 84.8	86.4 / 86.3	**86.9** / 86.2	84.1 / 83.8
Ref-L4 (val) [5]	78.6 / 83.6	85.8 / 87.5	86.0 / 89.0	89.5 / **90.0**	88.1 / 88.2	88.2 / 88.2
TallyQA [1]	75.4 / 75.0	80.3 / 79.9	81.4 / 81.5	**82.8** / 82.1	79.7 / 80.2	81.8 / 81.1
2D/3D Spatial Understanding						
CV-Bench [81]	80.3 / 76.4	86.1 / 86.1	88.0 / 86.3	**88.8** / 87.0	88.3 / 86.1	86.1 / 83.5
EmbSpatial-Bench [16]	72.1 / 63.8	79.5 / 75.7	81.3 / 76.1	**82.2** / 76.2	81.7 / 72.6	80.8 / 78.2
ERQA [77]	40.5 / 34.3	44.8 / 43.8	51.0 / 45.3	53.3 / 43.5	**55.3** / 45.5	43.8 / 41.3
MMSI-Bench [96]	27.8 / 29.2	30.2 / 28.3	33.8 / 33.0	**38.1** / 32.6	36.3 / 31.9	30.3 / 31.3
OmniSpatial [29]	43.3 / 40.8	49.1 / 51.0	54.1 / 50.9	56.2 / 53.2	**56.5** / 48.0	48.8 / 45.7
SpatialMQA [44]	40.7 / 37.6	45.0 / 43.6	51.5 / 47.3	**57.1** / 45.0	57.0 / 45.1	46.2 / 42.3

Continued on next page

Table 6 (continued)

Benchmark	WeMM-Ecom				Qwen3.5	GLM-4.6V-Flash
Size	**0.8B**	**2B**	**4B**	**9B**	**9B**	**9B**
Multi-Image Understanding						
BLINK [20]	57.9 / 52.8	64.2 / 63.8	66.1 / 65.8	68.5 / 68.1	**68.9** / 65.2	66.0 / 64.0
Mantis-Eval [31]	76.0 / 57.1	82.5 / 81.6	87.1 / 83.0	**88.0** / 81.6	87.6 / 82.0	83.4 / 77.4
MMIU [58]	54.1 / 46.0	61.3 / 58.5	65.8 / 63.6	**68.8** / 64.5	67.0 / 59.8	62.4 / 59.4
MuirBench [83]	52.1 / 35.1	59.6 / 59.8	65.5 / 66.2	71.4 / 63.0	67.3 / 52.4	**75.6** / 63.5
Instruction Following						
CFBench [106]	64.6 / 57.4	78.0 / 70.0	84.5 / 80.5	84.5 / 81.3	**84.7** / 83.0	81.8 / 74.5
IFBench [62]	29.7 / 28.0	45.3 / 33.3	56.3 / 32.7	68.0 / 50.7	**69.3** / 40.0	28.3 / 22.0
IFEval [108]	69.9 / 60.8	84.8 / 73.9	89.3 / 79.9	91.5 / 82.3	**92.4** / 84.5	63.4 / 57.3
MIA-Bench [64]	67.5 / 68.9	80.6 / 76.5	83.5 / 81.0	83.3 / **83.7**	83.1 / 82.7	81.1 / 77.0
MM-IFEval [15]	56.0 / 48.2	67.3 / 58.5	77.9 / 68.2	81.1 / 70.5	**81.3** / 64.8	66.1 / 57.7
Text-Centric Tasks						
C-Eval [26]	58.3 / 52.5	74.8 / 69.9	85.8 / 81.9	**88.2** / 86.3	88.0 / 86.7	84.7 / 77.6
GPQA (Diamond) [70]	35.4 / 36.4	60.1 / 54.6	71.7 / 67.2	75.8 / 75.8	78.3 / **78.8**	41.4 / 42.4
HMMT (Nov. 2025) [14]	10.0 / 3.3	27.5 / 33.3	70.8 / 68.3	80.0 / 75.8	**80.8** / 61.7	19.2 / 9.2
IMO-AnswerBench (v2) [51]	8.5 / 6.8	31.5 / 32.5	59.3 / 55.5	66.3 / 63.3	**66.8** / 48.3	25.8 / 20.8
MMLU-Pro [87]	49.6 / 47.6	69.2 / 66.2	80.0 / 77.6	**82.8** / 81.1	82.5 / 80.4	73.6 / 65.4
MMLU-Redux [21]	61.5 / 53.4	80.5 / 76.1	89.4 / 85.9	**92.0** / 89.7	91.5 / 90.1	87.0 / 69.9
SuperGPQA [18]	25.9 / 23.8	40.8 / 36.7	53.7 / 49.5	**58.7** / 55.2	58.6 / 55.3	45.9 / 34.8
Overall mean (64)	52.3 / 48.0	63.5 / 60.4	70.5 / 66.6	**74.2** / 69.7	73.4 / 67.6	63.3 / 57.9

5 Conclusion

We presented WeMM-Ecom, a family of five multimodal large language models ranging from a dense 0.8B model to a 35B-A3B mixture-of-experts model, together with a 28-benchmark e-commerce evaluation suite that spans eight business scenarios and 207,117 instances. Each model is trained with two stages of continued pre-training that raise the e-commerce share from 45.8% to 61.1%, and then post-trained with hybrid-thinking on-policy distillation, a single parameter-space merge with the post-trained Qwen3.5 backbone, and GRPO, yielding one checkpoint that serves both thinking and non-thinking modes. Across all ten size–mode comparisons, WeMM-Ecom improves on its Qwen3.5 baseline matched to size in e-commerce while matching or surpassing it on average across the 64 general benchmarks. We also show that WeMM-Ecom-35B-A3B attains the highest e-commerce average among evaluated open-source flagships with 5×–22× fewer total parameters, and WeMM-Ecom-9B surpasses the 3× larger Qwen3.8-27B. This indicates that domain specialization does not necessarily come at the expense of general capabilities, and scale does not gate the specialization benefit.

We release all model weights together with our data curation, training recipe, and benchmark construction so that this recipe can be examined and extended in future endeavors regarding domain-specialized models. In production, WeMM-Ecom already serves hundreds of workflows in WeChat E-Commerce, handling billion-scale requests per day across catalog management, multimodal product search, and livestream commerce. Future work will extend WeMM-Ecom to interactive shopping agents and study hardware-aware quantization for real-time edge deployment.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering Complex Counting Questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084, 2019.

[2] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, 2024.

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. PaliGemma: A Versatile 3B VLM for Transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[4] Haibin Chen, Kangtao Lv, Chengwei Hu, Yanshi Li, Yujin Yuan, Yancheng He, Xingyao Zhang, Langming Liu, Shilei Liu, Wenbo Su, et al. ChineseEcomQA: A Scalable E-commerce Concept Evaluation Benchmark for Large Language Models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5311–5321, 2025.

[5] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting Referring Expression Comprehension Evaluation in the Era of Large Multimodal Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 513–524. IEEE, 2025.

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are We on the Right Way for Evaluating Large Vision-Language Models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[7] Zeyu Chen, Guanghao Zhou, Min Yang, Qixiang Yin, Ziwang Zhao, Huanjin Yao, Pengjiu Xia, Cen Chen, and Minghui Qiu. Valley3: Scaling Omni Foundation Models for E-commerce. *arXiv preprint arXiv:2605.01278*, 2026.

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[9] Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. SimpleVQA: Multimodal Factuality Evaluation for Multimodal Large Language Models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4637–4646. IEEE, 2025.

[10] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. PaddleOCR 3.0 Technical Report. *arXiv preprint arXiv:2507.05595*, 2025.

[11] Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, et al. Aya Vision: Advancing the Frontier of Multilingual Multimodality. *arXiv preprint arXiv:2505.08751*, 2025.

[12] DeepSeek AI. DeepSeek-V4.1-Flash-1017: Wish-Driven Programming Made Real, 2026. URL https://api-docs.deepseek.com/news/news261017/.

[13] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104. IEEE, 2025.

[14] Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs. *arXiv preprint arXiv:2605.00674*, 2026.

[15] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. MM-IFEngine: Towards Multimodal Instruction Following. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1099–1109. IEEE, 2025.

[16] Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking Spatial Understanding for Embodied Tasks with Large Vision-Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, 2024.

[17] Wei Du, Shubham Toshniwal, Branislav Kisacanin, Sadegh Mahdavi, Ivan Moshkov, George Armstrong, Stephen Ge, Edgar Minasyan, Feng Chen, and Igor Gitman. Nemotron-Math: Efficient Long-Context Distillation of Mathematical Reasoning from Multi-Mode Supervision. *arXiv preprint arXiv:2512.15489*, 2025.

[18] Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, et al. SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines. *Advances in Neural Information Processing Systems*, 38, 2025.

[19] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning. *Advances in Neural Information Processing Systems*, 38, 2025.

[20] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[21] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are We Done with

MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, 2025.

[22] Jihao Gu, Yingyao Wang, Pi Bu, Chen Wang, Ziming Wang, Tengtao Song, Donglai Wei, Jiale Yuan, Yingxiu Zhao, Yancheng He, et al. See the World, Discover Knowledge: A Chinese Factuality Evaluation for Large Vision Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16422–16447, 2025.

[23] Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. CMMU: A Benchmark for Chinese Multi-Modal Multi-Type Question Understanding and Reasoning. *arXiv preprint arXiv:2401.14011*, 2024.

[24] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. *arXiv preprint arXiv:2507.01006*, 2025.

[25] Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. Step3-VL-10B Technical Report. *arXiv preprint arXiv:2601.09668*, 2026.

[26] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *Advances in Neural Information Processing Systems*, 36:62991–63010, 2023.

[27] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024.

[28] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations*, 2023.

[29] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards Comprehensive Spatial Reasoning Benchmark for Vision Language Models. In *International Conference on Learning Representations*, 2026.

[30] Ai Jian, Weijie Qiu, Xiaokun Wang, Peiyu Wang, Yunzhuo Hao, Jiangbo Pei, Yichen Wei, Yi Peng, and Xuchen Song. CSVQA: A Chinese Multimodal Benchmark for Evaluating STEM Reasoning Capabilities of VLMs. *arXiv preprint arXiv:2505.24120*, 2025.

[31] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved Multi-Image Instruction Tuning. *Transactions on Machine Learning Research*, 2024.

[32] Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, et al. Shopping MMLU: A Massive Multi-Task Online Shopping Benchmark for Large Language Models. *Advances in Neural Information Processing Systems*, 37:18062–18089, 2024.

[33] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An Optimizer for Hidden Layers in Neural Networks. Keller Jordan Blog, 2024. URL https://kellerjordan.github.io/posts/muon/.

[34] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth a Dozen Images. In *European Conference on Computer Vision*, pages 235–251. Springer, 2016.

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[36] Jovana Kondic, Pengyuan Li, Dhiraj Joshi, Isaac Sanchez, Ben Wiesel, Shafiq Abedin, Amit Alfassy, Eli Schwartz, Daniel Caraballo, Yagmur Gizem Cinar, et al. ChartNet: A Million-Scale, High-Quality Multimodal Dataset for Robust Chart Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[37] Jiale Li, Mingrui Wu, Zixiang Jin, Hao Chen, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. MIHBench: Benchmarking and Mitigating Multi-Image Hallucinations in Multimodal Large Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3143–3152, 2025.

[38] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded Language-Image Pre-Training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965. IEEE, 2022.

[39] Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. EcomGPT: Instruction-Tuning Large Language Models with Chain-of-Task Tasks for E-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18582–18590, 2024.

[40] Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. PerceptionBench: Evaluating Atomic Visual Perception in Multimodal Large Language Models. *arXiv preprint arXiv:2607.24957*, 2026.

[41] Xinyi Ling, Hanwen Du, Zhihui Zhu, and Xia Ning. EcomMMMU: Strategic Utilization of Visuals for Robust Multimodal E-commerce Models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 769–790, 2025.

[42] Liquid AI. Introducing LFM3: Every Modality for Specialized Intelligence. Liquid AI Blog, 2026. URL https://www.liquid.ai/blog/lfm3.

[43] Emmy Liu, Graham Neubig, and Chenyan Xiong. Midtraining Bridges Pretraining and Posttraining Distributions. *arXiv preprint arXiv:2510.14865*, 2025.

[44] Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can Multimodal Large Language Models Understand Spatial Relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632, 2025.

[45] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is Your Multi-Modal Model an All-Around Player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2024.

[46] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *Science China Information Sciences*, 67(12):220102, 2024.

[47] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

[48] Chenji Lu, Zhuo Chen, Hui Zhao, Zhenyi Wang, Pengjie Wang, Chuan Yu, and Jian Xu. RAIR: A Rule-Aware Benchmark Uniting Challenging Long-Tail and Visual Salience Subset for E-commerce Relevance Assessment. *arXiv preprint arXiv:2512.24943*, 2025.

[49] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations*, 2024.

[50] Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, et al. FinMME: Benchmark Dataset for Financial Multi-Modal Reasoning Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29465–29489, 2025.

[51] Minh-Thang Luong, Dawsen Hwang, Hoang H Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, et al. Towards Robust Mathematical Reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35418–35442, 2025.

[52] Shirong Ma, Shen Huang, Shulin Huang, Xiaobin Wang, Yangning Li, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. EcomGPT-CT: Continual Pre-Training of E-commerce Large Language Models with Semi-Structured Data. *arXiv preprint arXiv:2312.15696*, 2023.

[53] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

[54] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022.

[55] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, 2025.

[56] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A Dataset for VQA on Document Images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208. IEEE, 2021.

[57] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591. IEEE, 2022.

[58] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. MMIU: Multimodal Multi-Image Understanding for Evaluating Large Vision-Language Models. In *International Conference on Learning Representations*, 2025.

[59] OpenCompass. CCBench: Chinese Cultural Benchmark, 2023. URL https://github.com/open-compass/MMBench#news. Dataset release in the official MMBench repository.

[60] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24838–24848. IEEE, 2025.

[61] Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. eCeLLM: Generalizing Large Language Models for E-commerce from Large-Scale, High-Quality Instruction Data. *arXiv preprint arXiv:2402.08831*, 2024.

[62] Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hanna Hajishirzi. Generalizing Verifiable Instruction Following. *Advances in Neural Information Processing Systems*, 38, 2025.

[63] Haonan Qi, Jin Cao, Yongqi Zhang, Xintong Wang, Weidong Tang, Bin Chen, Chengfu Huo, Haojun Pan, Hengyu You, Jing Li, et al. IndustryBench-MIPU: Benchmarking Multi-Image Attribute Value Extraction for Industrial Products. *arXiv preprint arXiv:2606.14383*, 2026.

[64] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. MIA-Bench: Towards Better Instruction Following Evaluation of Multimodal LLMs. In *International Conference on Learning Representations*, 2025.

[65] Zongming Qiao, Ziyue Hua, Zirui Ou, Zihao Yue, Zihan Jiang, Zhuo Huang, Zhiyang Chen, Zhixian Zheng, Zhipeng Xu, Zhengrui Ma, et al. MiMo-V2.6: Scaling Reinforcement Learning Towards Self-Improvement. *arXiv preprint arXiv:2610.11959*, 2026.

[66] Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the Design of Qwen3.8-Next Architecture: Evaluation, Efficiency, and Training Stability. *arXiv preprint arXiv:2608.30320*, 2026.

[67] Qwen Team. Qwen3.6-35B-A3B Model Card. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B.

[68] Qwen Team. Qwen3.8-27B Model Card. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

[69] Qwen Team. Qwen3.5: Towards Native Multimodal Agents. Official release blog and model cards, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

[70] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[71] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. ZeroBench: An Impossible Visual Benchmark for Contemporary Large Multimodal Models. *arXiv preprint arXiv:2502.09696*, 2025.

[72] Andrew Sellergren, Sahar Kazemzadeh, Fereshteh Mahvar, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Chufan Gao, Ronnachai Jaroensri, Shawn Xu, Fayaz Jamil, et al. An Open Vision-Language Model for Diverse Medical Applications. *Nature Medicine*, 32(10):3695–3704, 2026.

[73] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jia Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019.

[74] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

[75] Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. VisualPuzzles: Decoupling Multimodal Reasoning Evaluation from Domain Knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

[76] Jayant Sravan Tamarapalli, Rynaa Grover, Nilay Pande, and Sahiti Yerramilli. CountQA: How Well Do MLLMs Count in the Wild? *arXiv preprint arXiv:2508.06585*, 2025.

[77] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the Physical World. *arXiv preprint arXiv:2503.20020*, 2025.

[78] Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026.

[79] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-VL Technical Report. *arXiv preprint arXiv:2504.07491*, 2025.

[80] Tencent. Tencent Launches HY4, Joining the Ranks of the World's Leading Open-Source AI Models. Official model release announcement, November 2026. URL https://www.tencent.com/tencent-launches-hy4-joining-the-ranks-of-the-worlds-leading-open-source-ai-models/.

[81] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[82] An-Lan Wang, Jingqun Tang, Lei Liao, Hao Feng, Qi Liu, Xiang Fei, Jinghui Lu, Han Wang, Hao Liu, Yuliang Liu, et al. WildDoc: How Far Are We from Achieving Comprehensive and Robust Document Understanding in the Wild? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23002–23012, 2025.

[83] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. MuirBench: A Comprehensive Benchmark for Robust Multi-Image Understanding. In *International Conference on Learning Representations*, 2025.

[84] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[85] Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, et al. LocateAnything: Fast and High-Quality Vision-Language Grounding and Detection. *arXiv preprint arXiv:2605.27365*, 2026.

[86] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[87] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[88] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

[89] Zichen Wen, Boxue Yang, Shuang Chen, Yaojie Zhang, Yuhang Han, Junlong Ke, Cong Wang, Yicheng Fu, Jiawang Zhao, Jiangchao Yao, et al. Innovator-VL: A Multimodal Large Language Model for Scientific Discovery. *arXiv preprint arXiv:2601.19325*, 2026.

[90] Junyu Wu, Weiming Chang, Xiaotao Liu, Guanyou He, Tingfeng Xian, Haoqiang Hong, Boqi Chen, Hongtao Tian, Tao Yang, Yunsheng Shi, et al. WeChat-YATT: A Scalable, Simple, Efficient, and Production Ready Training Library. *arXiv preprint arXiv:2508.07970*, 2025.

[91] Ziheng Wu, Zhenghao Chen, Ruipu Luo, Can Zhang, Yuan Gao, Zhentao He, Xian Wang, Haoran Lin, and Minghui Qiu. Valley2: Exploring Multimodal Models with Scalable Vision-Language Design. *arXiv preprint arXiv:2501.05901*, 2025.

[92] xAI. Grok-1.5 Vision Preview, 2024. URL https://x.ai/news/grok-1.5v. RealWorldQA benchmark release.

[93] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[94] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. VisuLogic: A Benchmark for Evaluating Visual Reasoning in Multi-Modal Large Language Models. In *International Conference on Learning Representations*, 2026.

[95] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

[96] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence. In *International Conference on Learning Representations*, 2026.

[97] Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, et al. CC-OCR: A Comprehensive and Challenging OCR Benchmark for Evaluating Large Multimodal Models in Literacy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21744–21754. IEEE, 2025.

[98] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling Context in Referring Expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[99] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and

Training Recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11704–11715, 2026.

[100] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *International Conference on Machine Learning*, 2024.

[101] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[102] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. MMMU-Pro: A More Robust Multi-Discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, 2025.

[103] Z.ai. GLM-5.3-Flash Model Card, 2026. URL https://huggingface.co/zai-org/GLM-5.3-Flash.

[104] Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, et al. Can MLLMs Understand the Deep Implication Behind Chinese Images? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402, 2025.

[105] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. CMMMU: A Chinese Massive Multi-Discipline Multimodal Understanding Benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

[106] Tao Zhang, Chenglin Zhu, Yanjun Shen, Wenjing Luo, Yan Zhang, Hao Liang, Fan Yang, Mingan Lin, Yujing Qiao, Weipeng Chen, et al. CFBench: A Comprehensive Constraints-Following Benchmark for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32926–32944, 2025.

[107] Yi Zhang, Bolin Ni, Xin-Sheng Chen, Hengrui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Winston Hu, Meng-Hao Guo, and Shi-Min Hu. Bee: A High-Quality Corpus and Full-Stack Suite to Unlock Advanced Fully Open MLLMs. In *International Conference on Learning Representations*, 2026.

[108] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*, 2023.

[109] Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. WeMM-Embedding: WeChat Multi-Modal Embedding Technical Report. *arXiv preprint arXiv:2608.24053*, 2026.

[110] Runjie Zhou, Youbo Shao, Haoyu Lu, Bowei Xing, Tongtong Bai, Yujie Chen, Jie Zhao, Lin Sui, Haotian Yao, Zijia Zhao, et al. WorldVQA: Measuring Atomic World Knowledge in Multimodal Large Language Models. *arXiv preprint arXiv:2602.02537*, 2026.

[111] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. In *International Conference on Learning Representations*, 2025.

A Training and Data Details

A.1 Post-Training Hyperparameters and Prompt Curation

Distillation and RL sampling. During OPD, we sample two responses per prompt at temperature 1.0 and top-p 1.0, with a maximum generation length of 24,576 tokens within a 32,768-token sequence limit. The distillation loss weight is set to 1.0 without an additional reference-policy penalty; compact students (0.8B, 2B, and 4B) initialize from their CPT checkpoints with zero weight decay. During RL, we sample $G = 16$ responses per prompt within a 65,536-token context limit using weight decay 0.01, temperature 1.0, and top-p 0.9 by default (supplemented with top-k 20 and repetition penalty 1.05 on instruction-following tasks). All RL runs initialize both the actor and frozen reference policy from the equally weighted ($\alpha = 0.5$) post-OPD merged checkpoint.

Post-training prompt screening. To construct the OPD prompt pool, both the teacher and student models generate eight candidate responses per prompt; we retain prompts where the teacher achieves a strictly higher mean quality score than the student. For the RL pool, the corresponding post-OPD merged model generates eight thinking-mode rollouts per candidate prompt. We retain prompts with positive reward variance and a mean task reward of at most 0.95, and additionally require a strictly positive mean reward on mathematical and logical reasoning tasks.

Mode assignment. Operating modes are assigned deterministically by hashing each prompt identifier so that all rollouts for a given prompt share the same mode. Thinking mode accounts for 400,229 prompts (80.05%) in the 499,998-prompt OPD pool and 24,823 prompts (80.14%) in the 30,976-prompt RL pool.

A.2 Reward Design and RL Optimization

In Equation 4, group standard deviations are computed using Bessel's correction with stabilizer $\epsilon_A = 10^{-4}$ (groups with identical rewards receive zero advantage). In Equation 5, the token-level KL penalty is computed via the Schulman approximator $K_{i,t} = \min\big(10, \exp(d_{i,t}) - d_{i,t} - 1\big)$ with $d_{i,t} = \mathrm{clip}\big(\log \pi_{\mathrm{ref}}(y_{i,t} \mid h_{i,t}) - \ell_{i,t}, -20, 20\big)$ and KL coefficient 0.01.

E-commerce and general task rewards. Task rewards combine deterministic rule verifiers with model-based judgments from Qwen3.5-122B-A10B. Product and SKU retrieval tasks assign unit reward for predicting the correct equivalence label in the required schema. For localization and detection tasks, let N denote the total number of predicted and ground-truth boxes and $\mathcal{P}$ denote the optimal one-to-one matching within each image. For $N > 0$, the detection reward is

$$R_{\mathrm{det}} = \frac{2}{N} \sum_{(p,g)\in\mathcal{P}} \mathrm{IoU}(p, g)\, w(p, g), \tag{6}$$

where $w(p, g) = 1$ for livestream and multi-image detection, and $w(p, g) = 0.7 + 0.3\, s(p, g)$ for target-product detection ($s = 1$ for exact class match, 0.5 for partial/full product class confusion, and 0 otherwise). For instruction-following tasks, full constraint satisfaction receives reward 1.0, while partial compliance receives 0.49 times the mean constraint satisfaction score evaluated on the visible final answer after stripping the reasoning block.

Format validation. Rollouts that violate the required output schema or hit the generation cap before completion receive zero task reward; in thinking mode, responses must contain a well-formed reasoning block followed by a non-empty final answer. If verifier evaluation fails on a prompt group, the task-reward term is masked out and only the reference KL penalty is applied.

B Evaluation Benchmarks and Protocols

Decoding and semantic adjudication. For WeMM-Ecom, thinking mode uses temperature 1.0, top-p 0.95, and top-k 20, paired with presence penalty 2.0 and repetition penalty 1.05 on reasoning-intensive benchmarks

and presence penalty 1.5 with repetition penalty 1.0 on the remaining tasks. In non-thinking mode, standard benchmarks use temperature 0.7 and top-p 0.8, while open-ended generation subsets use temperature 1.0, top-p 1.0, and top-k 40. All models use a default maximum generation limit of 65,536 tokens (with greedy decoding at 32,768 tokens for OmniDocBench). Baseline models are evaluated under their recommended or task-calibrated decoding parameters. Model-based evaluation adjudication uses Qwen3-30B-A3B-Instruct-2507 [95].

Table 7 E-commerce evaluation sets and scoring metrics.

Benchmark	Instances	Input and target	Primary metric
Product Knowledge			
E-commerce Knowledge	4,048	Text questions about products, brands, and industry knowledge	Judged accuracy
Entity Recognition	550	Product image to product name, brand, and model	Acceptability
Product Attributes			
Attribute Selection	3,020	Product image and candidate values to one attribute value	Accuracy
Attribute Extraction	14,534	Listing images, product fields, and options to attribute values	Reference-attribute accuracy
Brand Attribution	6,122	Listing images and fields, with optional brand candidates, to product brand	Accuracy
Brand Extraction	3,000	Listing images and product context to an extracted brand name	Accuracy
Product Moderation			
Illegal Goods	3,496	Listing evidence and rules to an illegal-goods decision	Accuracy
Sales Restrictions	1,050	Listing evidence and rules to a restricted-sales decision	Accuracy
Image-Title Mismatch	1,897	Product title and images to the set of unrelated image indices	Exact-set accuracy
Product OCR			
Product OCR	19,962	Single, multiple, or long product images to an answer about visible text	Mean judged accuracy
Product Retrieval			
Region-level Retrieval	15,335	Two product crops to an ordinal similarity judgment	Normalized Spearman
Product-level Retrieval	2,981	Two product images and listing fields to a same-product judgment	Average precision
SKU-level Retrieval	3,268	Two textual SKU descriptions to a variant-equivalence decision	Match F1
Product Grounding			
Chart Detection	1,456	Listing image to chart and table boxes	Mean class F1
Product Detection	5,000	Product image to boxes for the target product	Box F1
Logo Detection	1,500	Product image to brand-logo boxes	Box F1
Qualification Detection	997	Product image to typed certification-mark boxes	Mean class F1
Dimension Detection	994	Product image to dimension-annotation boxes	Box F1
Livestream Detection	1,992	Livestream frame to boxes for the advertised product	Box F1
E-commerce Reasoning			
Chart Reasoning	2,000	Rendered SKU, price, attribute, or constraint table to a derived answer	Mean accuracy
Category Reasoning	500	Product image and options to the requested category	Accuracy

Continued on next page

Table 7 (continued)

Benchmark	Instances	Input and target	Primary metric
Product Counting	500	Product arrangement and selection constraints to an item count	Accuracy
Spatial Reasoning	500	Product arrangement and relative-position constraints to a target item	Accuracy
E-commerce (Open)			
ShoppingMMLU	20,570	Text tasks spanning shopping knowledge and e-commerce operations	Mean task score
RAIR	48,835	Query, item fields, and available images to a relevance grade	Macro F1
ChineseEcomQA	1,800	Chinese e-commerce questions and task instructions to answers or labels	Mean task score
EcomMMMU (GTS)	36,736	Main product image and text to outputs for eight GTS tasks	GTS mean score
IndustryBench-MIPU	4,474	Multiple industrial-product images and a request to attribute values	Micro F1

Each instance may contain several images, targets, or products.

Benchmark construction and answer adjudication. Table 7 summarizes the 28 e-commerce benchmarks. Product OCR averages accuracies equally across single-image (4,970 instances), multi-image (6,494), and long-image (8,498) subsets. Chart Reasoning averages four 500-question subsets covering SKU charts, price tables, attribute tables, and constraint tables; questions are synthesized under category-specific templates with programmatically computed reference answers and cross-checked by frontier LLMs for reasoning consistency. During evaluation, rule-based matchers first check typed answers and taxonomy aliases before invoking the LLM judge for semantic equivalence. Entity Recognition credits full or partial identifications, Brand tasks allow semantic alias matching, and Image-Title Mismatch requires exact set agreement.

Grounding evaluation protocol. All six grounding benchmarks use an IoU threshold of 0.5 with coordinates normalized to $[0, 1000]$. Chart Detection and Qualification Detection report macro-averaged class F1, whereas the remaining sets report box F1 (with Product Detection using greedy one-to-one matching). In Chart Detection, prompts request JSON objects with `bbox_2d` and `category`; to ensure schema compatibility across model families, the evaluation parser also accepts class labels output under the `label` key when `category` is absent.

Retrieval and public benchmark protocols. In Region-level Retrieval, Spearman rank correlation ρ is linearly rescaled to $50(\rho + 1)$; for Region-level and Product-level Retrieval, non-thinking mode scores token probabilities whereas thinking mode parses generated labels. Among the public benchmarks, EcomMMMU [41] is evaluated on its labeled general test set (GTS) using the main product image and reported as the unweighted mean across its eight subtasks, while IndustryBench-MIPU [63] uses multi-image product inputs and reports attribute-extraction micro F1.