Grounded Multimodal In-Context Learning for Product Weight Estimation at Scale in E-commerce

Bhavuk Singhal
Meesho, India
bhavuk.singhal@meesho.com

Arsh Keshari
Meesho, India
arsh.keshari@meesho.com

Ravindra Kumar Yadav
Meesho, India
ravindra@meesho.com

Abstract

Accurately inferring implicit physical attributes of products, such as weight, is critical for large-scale e-commerce logistics but challenging due to sparse or unreliable textual metadata and high visual variability. We formulate weight estimation as a grounded multimodal reasoning problem and investigate whether large vision-language models (LVLMs) can infer discretized weight buckets through in-context learning (ICL) over product images and descriptions. We introduce a scalable inference framework that conditions predictions on automatically retrieved, category-specific exemplars and propose a distribution-calibrated retrieval strategy that aligns few-shot contexts with the empirical weight distribution of each product subcategory. Offline evaluations shows that across 14 high-variance product categories, our approach significantly outperforms strong multimodal KNN baselines in both exact-match accuracy and near-bucket reliability. Deployed in production on a large e-commerce platform, our system processes millions of listings daily and reduces shipping-related revenue leakage by $\sim$22%, demonstrating that multimodal ICL can serve as a practical and cost-effective alternative to manual or hardware-based verification.

1 Introduction

In large-scale e-commerce platforms, many operational decisions depend on estimating implicit physical attributes of products, such as weight, that are rarely stated explicitly in textual metadata. Inferring these attributes from noisy product listings requires grounding language descriptions with visual cues including size, material, and packaging density. Recent advances in large vision-language models (LVLMs) enable multimodal reasoning over images and text through in-context learning, offering a promising alternative to manual verification or specialized hardware. However, it remains unclear whether such models can reliably perform fine-grained physical inference at the scale and variability encountered in real-world commerce.

Weight estimation presents a particularly challenging setting. At the most granular level of product taxonomy, Sub-Sub-Categories (SSCats), visually similar items often exhibit large variations in mass due to differences in material composition and construction. For instance, products in the same "Spice Rack" (Figures 1a, 1b, 1c) or "Mops with Bucket" (Figures 1d, 1e, 1f) category may vary by several kilograms despite looking nearly identical. This high intra-category variance creates a long-tailed distribution where simple heuristics or average-based predictions fail to generalize.

Existing solutions primarily rely on manual audits, hardware-based weighing (Shiprocket, 2025), or unimodal machine learning models. Although hardware methods are accurate, they are costly and hard to scale across millions of listings. Traditional models like KNN or Random Forests (Bain & Company, 2025; Verma et al., 2025) depend on shallow text features and historical averages, lacking the multimodal reasoning needed to infer material density or volume. As a result, they often produce consistent under- or over-estimates, leading to financial losses and operational inefficiencies.

This work investigates multimodal ICL with LVLMs as a scalable framework for grounded physical attribute inference. Compared to fine-tuning approaches, ICL enables rapid onboarding of new product categories without additional retraining or labeled data, which is critical for continuously evolving e-commerce catalogs. Our contributions are summarized below.

- We formulate product weight estimation as an implicit physical attribute inference problem and show that LVLMs can perform grounded multimodal reasoning over images and textual metadata through few-shot ICL.
- We propose an automated, category-aware re-

(a) 2 kg (b) 4 kg (c) 7 kg

(d) 2.5 kg (e) 8 kg (f) 11 kg

Figure 1: Products images listed on e-commerce platform.

trieval strategy that aligns few-shot exemplars with the empirical weight distribution of each product sub-category. We demonstrate that this calibration substantially improves accuracy and reduces misestimation compared to other methods.

- We conduct systematic studies on exemplar quantity, retrieval strategy, and model backbones, revealing accuracy saturation effects and non-linear cost–latency trade-offs that inform practical deployment of LVLM-based inference at scale.

- Deployed in production across four high-variance categories and millions of daily listings, our approach improves exact bucket accuracy by 34% and reduces shipping-related revenue leakage by 22%, demonstrating the real-world viability of multimodal ICL for logistics optimization.

2 Related Work

Estimating physical product attributes from catalog metadata sits at the intersection of the following systems:

Attribute Value Extraction (AVE). Early AVE methods relied on sequence labeling and structured prediction (e.g., Bi-LSTMs, CRFs) to extract explicit textual signals (Zheng et al., 2018; Zhu et al., 2020; Xu et al., 2019; Wang et al., 2020; Gong and Eldardiry, 2024). However, these unimodal approaches struggle to infer implicit physical properties like weight or volume. Recent Large Vision-Language Models (LVLMs)—including LLaVA (Liu et al., 2023a), InstructBLIP (Dai et al., 2023), and BLIP-2 (Li et al., 2023)—enable joint reasoning over images and text for latent attribute inference. While EIVEN (Zou et al., 2024) utilizes generative estimation, our work focuses on scalable, high-throughput multimodal in-context learning for industrial settings using a native LVLM.

ICL and Retrieval. ICL allows models to adapt to new tasks via few-shot demonstrations (Brown et al., 2020; Dong et al., 2024; Von Oswald et al., 2023), but performance is highly sensitive to exemplar selection and ordering (Lu et al., 2022; Liu et al., 2022). This is particularly challenging for high-cardinality tasks like weight bucket prediction, where visual and material cues vary significantly (Milios et al., 2023). While prior strategies use active learning or retrieval to improve context quality (Rubin et al., 2022; Ram et al., 2023), they are often computationally expensive. We propose a lightweight, distribution-calibrated retrieval strategy that aligns exemplars with category-specific distributions, enhancing calibration and efficiency for large-scale deployment.

Industrial Logistics Systems. Inaccurate weight and dimension estimates cause operational inefficiencies and revenue leakage (Eshopbox, 2025; Redseer, 2024). While hardware-based systems and manual audits are precise, they are costly and difficult to scale (Shiprocket, 2025). Current software baselines (e.g., KNN or tree-based models) rely on historical averages or shallow embeddings, failing to reason about material density or structural cues (Kasemrat and Kraiwanit, 2024; Ziabari, 2025). Our work bridges this gap by using LVLM-based multimodal in-context inference as a scalable, cost-effective, and efficient alternative to traditional verification methods.

3 Methodology

We study the problem of inferring a product's chargeable weight from multimodal inputs $x_i = (I_i, T_i)$, where weight is rarely stated explicitly and must instead be inferred from visual and linguistic cues such as size, material, and packaging. Rather than predicting a continuous value, we discretize the label space into K ordered weight buckets $\mathcal{B} = \{b_1, \ldots, b_K\}$, since fine-grained visual estimates (e.g., 1.1 kg vs. 1.2 kg) are noisy and operationally unstable. This bucketed formulation yields a more robust and billing-aligned decision boundary. Formally, we define a mapping function $f : (I, T, \mathcal{E}_C) \rightarrow \mathcal{B}$ conditioned on a set of category-specific exemplars $\mathcal{E}_C$.

3.1 Distribution-Calibrated Exemplar Retrieval

Effective in-context learning critically depends on the quality and distribution of the provided exemplars. Manually curating representative products for every new SSCat is impractical at scale. We therefore introduce an automated retrieval strategy that selects reliable, category-specific exemplars from historical data to construct informative few-shot contexts. Our retrieval procedure is guided by two principles: (i) identifying high-confidence instances with trustworthy labels, and (ii) matching the empirical weight distribution of the target category to avoid contextual bias.

Implicit Verification via Order Velocity. In the absence of universal physical verification, we use sales velocity as a proxy for label reliability. We define a candidate pool S_{valid} consisting of products whose delivered orders $O_{\text{delivered}}$ exceed a threshold θ. High-velocity items are less likely to contain incorrect declared weights, as frequent shipments would otherwise trigger operational disputes. We therefore treat their historical weights as reliable soft labels for exemplar construction.

Distribution-Calibrated Sampling. Few-shot performance can degrade when the exemplar distribution is misaligned with the true category distribution. To mitigate this effect, we estimate the empirical weight distribution $P(b)$ over buckets $b \in \mathcal{B}$ within S_{valid}, and allocate the context budget N_{shot} proportionally according to $P(b)$. This ensures that the retrieved exemplars reflect the natural frequency of light and heavy items, improving calibration of the model's predictions.

The complete retrieval procedure is summarized in Algorithm 1.

Algorithm 1: Exemplar Retrieval

Input:
Historical product set H;
weight buckets $B = \{b_1, \ldots, b_K\}$;
minimum delivered-order threshold θ;
number of exemplars N_{shot}
Output: Retrieved exemplar set $\mathcal{E}_C$
$S_{\text{valid}} \leftarrow \emptyset$;
foreach $p \in H$ **do**
  if $O_{delivered}(p) \geq \theta$ **then**
    Assign p to bucket $b_k \in B$ based on weight y_p;
    $S_{\text{valid}} \leftarrow S_{\text{valid}} \cup \{(I_p, T_p, b_k)\}$;
if $|S_{valid}| = 0$ **then**
  return fallback exemplar set;
foreach $b_k \in B$ **do**
  $c_k \leftarrow |\{e \in S_{\text{valid}} : e.\text{bucket} = b_k\}|$;
$n \leftarrow |S_{\text{valid}}|$;
foreach $b_k \in B$ **do**
  $P(b_k) \leftarrow \frac{c_k}{n}$;
  $n_k \leftarrow \lfloor P(b_k) \times N_{\text{shot}} \rfloor$;
Adjust $\{n_k\}$ such that $\sum_k n_k = N_{\text{shot}}$;
$\mathcal{E}_C \leftarrow \emptyset$;
foreach $b_k \in B$ **do**
  $S_k \leftarrow \{e \in S_{\text{valid}} : e.\text{bucket} = b_k\}$;
  Sample $\min(n_k, |S_k|)$ exemplars uniformly from S_k;
  $\mathcal{E}_C \leftarrow \mathcal{E}_C \cup S_k^{\text{sampled}}$;
if $|\mathcal{E}_C| < N_{shot}$ **then**
  Sample additional exemplars from $S_{\text{valid}} \setminus \mathcal{E}_C$;
  Add samples to $\mathcal{E}_C$ until $|\mathcal{E}_C| = N_{\text{shot}}$;
return E_C;

3.2 Multimodal In-Context Inference with LVLMs

We leverage Gemini 2.5 Flash (Comanici et al., 2025), a large vision-language model (LVLM), for native multimodal reasoning over interleaved image–text inputs. LVLMs enable joint grounding of visual cues (e.g., size, texture, and packaging density) with linguistic metadata, making them well-suited for inferring implicit physical attributes such as weight. Given a query product $x_q = (I_q, T_q)$, we construct a single-turn in-context prompt S consisting of a task instruction followed by the retrieved exemplars and the query:

$$S = [\ \text{Inst}, (I_1, T_1, y_1), \ldots, (I_{N_{\text{shot}}}, T_{N_{\text{shot}}}, y_{N_{\text{shot}}}), (I_q, T_q)\ ] \quad (1)$$

This formulation encourages the model to perform comparative reasoning by aligning the query against labeled anchors within the same category. The model then predicts a discrete bucket $\hat{y} \in \mathcal{B}$ conditioned on the provided context. Additional details on prompt structure and generation settings are provided in Appendix B.

3.3 Multimodal KNN Baseline

To establish a strong non-generative baseline, we compare against nearest-neighbor retrieval in a

(a) KNN-11kg, Ours-20kg (b) KNN-6kg, Ours-10kg (c) KNN-7kg, Ours-12.5kg (d) KNN-6kg, Ours-10.5kg

Figure 2: Qualitative Analysis

shared multimodal embedding space. We use the FLAVA model (Singh et al., 2021) to compute joint image–text embeddings that align visual and textual features into a unified representation. All catalog embeddings are indexed using FAISS (Douze et al., 2025) for efficient similarity search. For a query x_q, we retrieve its nearest neighbors and assign a weight bucket based on two strategies: **Top-1**, which directly adopts the closest neighbor's label, and **Top-5 Avg**, which averages the labels of the five nearest neighbors before discretization. This baseline isolates the effect of embedding similarity without in-context reasoning, providing a direct comparison to the proposed LVLM-based approach.

4 Empirical Study

4.1 Experimental Setup

Data Construction. We construct an 18-month evaluation corpus across 14 sub-categories (Table 4). To prevent temporal leakage, we use a strict chronological split: the first 12 months form the retrieval pool, while the final 6 months are reserved for evaluation. Ground-truth weights are manually verified by experts at fulfillment centers to ensure high labeling fidelity.

Evaluation Framework. The target weight range spans 500 g to 25 kg and is discretized into $K = 50$ uniform buckets $\mathcal{B} = \{b_1, \ldots, b_K\}$ of 500 g increments. We evaluate performance using Accuracy (Acc) for exact bucket matches, and Acc@$\pm$1 and @$\pm$2 to measure near-miss robustness within one and two neighboring buckets. Additionally, we track the Abstention Rate (AR) for queries where the model declines to predict due to low confidence or insufficient context. Finally, Revenue Leakage (RevL (% Δ)) measures the relative percentage difference in shipping charges between predicted and verified weights, quantifying the financial impact of misestimation in deployment terms.

4.2 Proposed Approach vs. KNN

Quantitative Results. As shown in Table 1, our LVLM-based in-context approach consistently outperforms multimodal KNN baselines. While KNN often fails due to high abstention rates on diverse products, our method ensures robust predictions and superior calibration across all accuracy metrics (Exact Match, Acc@$\pm$1, and Acc@$\pm$2). By leveraging multimodal reasoning over simple embedding similarity, we significantly reduce average RevL, demonstrating the practical utility of our approach for industrial-scale weight inference.

Qualitative Analysis. Case studies illustrate how our LVLM outperforms embedding-based retrieval by integrating fine-grained visual and textual cues. While KNN relies on global similarity, the LVLM reasons about material density and scale: it detects printed "20KG" text in Compost (Fig. 2a), identifies heavy-duty Kitchen Storage via metallic textures (Fig. 2b), and distinguishes industrial steel from plastic in Mops (Fig. 2c). In Spice Racks (Fig. 2d), reasoning about glass and metal density prevents underestimation. See Appendix C for further examples (with textual descriptions).

4.3 Impact of Input Modality

To quantify modality contributions, we conducted an ablation study across three settings: *Text-only* (titles and descriptions), *Image-only*, and Multimodal (our full method). As shown in Table 1, removing either modality significantly degrades performance. Text-only fails when weight depends on visual cues

Table 1: Performance comparison

Model Strategy	Acc(%)	Acc@±1(%)	Acc@±2(%)	AR(%)	RevL (% Δ)
KNN (Top-1)	36.42	39.86	45.58	28.41	-5.6%
KNN (Top-5 Avg)	42.28	43.48	51.37	23.92	-8.4%
Ours	**73.81**	**76.01**	**85.43**	**0.1**	**-42.3%**
w Text-only	43.12	46.55	54.28	0.6	-14.2%
w Image-only	51.45	55.20	62.10	0.9	-21.8%
w Random Sampling	65.50	70.24	77.59	0.2	-29.8%

like material density whereas Image-only struggles without textual metadata or OCR signals. The substantial improvement of the multimodal approach confirms that accurate weight estimation requires grounded reasoning across both signal sources.

4.4 Impact of Exemplar Quantity (N_{shot})

We next analyze how the number of in-context exemplars affects multimodal ICL performance. Varying N_{shot} controls the amount of category-specific evidence available to the LVLM, and therefore directly influences both predictive accuracy and inference cost. We evaluate N_{shot} from 5 to 100 while keeping all other components fixed.

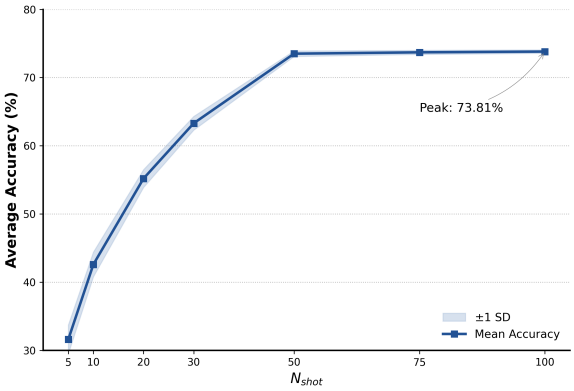

Figure 3: Average accuracy vs. N_{shot}

Accuracy Saturation. Figure 3 illustrates a logarithmic accuracy improvement as N_{shot} increases. Performance gains are rapid for small context sizes, confirming the efficiency of few-shot multimodal reasoning. However, gains diminish beyond $N_{shot} = 50$, with additional exemplars yielding less than 0.5% improvement. This saturation suggests that once the context sufficiently covers a category's weight distribution, further examples provide limited additional signal.

Cost and Latency Trade-offs. Increasing N_{shot} expands prompt length, directly impacting inference cost and latency. Figure 4 shows a non-linear spike in both metrics around 55–60 shots due to LVLM token limits. While larger contexts offer marginal accuracy gains, they incur disproportionate computational overhead. Balancing performance and efficiency, we select $N_{shot} = 50$ as the optimal operating point for our experiments.

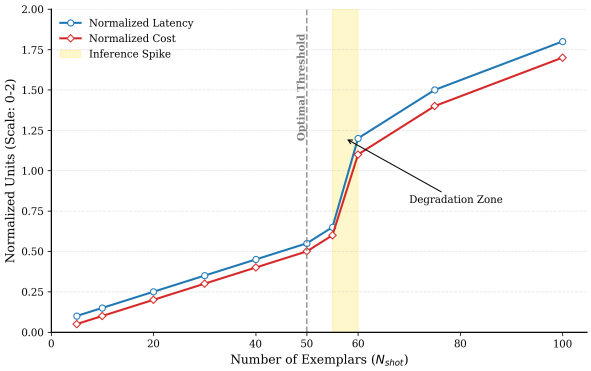

Figure 4: Inference cost and latency vs. N_{shot}

4.5 Impact of Retrieval Strategy

We evaluate how exemplar selection impacts multimodal ICL by comparing our distribution-calibrated retrieval (Algorithm 1) against a random sampling baseline. As shown in Table 1, distribution-calibrated retrieval consistently improves accuracy and reduces RevL. While random sampling often introduces bias by over-representing rare or extreme-weight items, aligning exemplars with the true category distribution provides a more faithful few-shot context. These results confirm that exemplar distribution—not just the number of shots—is critical for stable model calibration.

4.6 LVLM Backbone Benchmarking

We evaluate how different LVLM backbones affect multimodal ICL performance, balancing inference quality against deployment cost. We benchmarked several proprietary and open-source models supporting interleaved prompts, including Gemini,

GPT-4o (Hurst et al., 2024), LLaVA-v1.6-34B (Liu et al., 2023b), and Qwen2-VL-72B (Wang et al., 2024). All models were tested under the same few-shot settings for classification accuracy and normalized inference cost (NC). Table 2 shows that while frontier models like Gemini 1.5 Pro and GPT-4o achieve marginally higher accuracy, they incur substantially higher costs. Conversely, smaller open-source models exhibit lower accuracy and reduced long-context robustness. Gemini 2.5 Flash provides the optimal accuracy–cost trade-off, matching the performance of larger models with significantly lower latency and expense. Consequently, we adopted it for production deployment.

Table 2: LVLM Backbone Benchmarking

Model	Acc(%)	NC
Gemini 2.5 Flash	73.81	∼1.0x
Gemini 1.5 Pro	74.23	∼2.57x
GPT-4o	73.89	∼5.15x
GPT-4o mini	66.45	∼0.81x
LLaVA	56.24	∼0.15x
Qwen2	59.13	∼0.13x

5 Production Deployment

To evaluate the practicality of our approach under real operational conditions, we deployed the proposed LVLM-based inference system in a live production environment across four high-variance SSCats (*Compost & Fertilizers*, *Kitchen Storage*, *Mops with Bucket*, and *Spice Racks*). The deployment processes millions of listings and directly replaces manual or heuristic weight assignment.

5.1 Results Summary

Table 3 reports performance using the same metrics defined in Section 4.1, enabling direct comparison with offline evaluation. In addition, we report the *Undercharging Rate (UC)*, defined as the percentage of orders where the predicted bucket is lower than the verified ground-truth bucket, and *UC (2+)*, which counts underestimation by two or more buckets. The deployed system improves weight bucket accuracy while substantially reducing systematic underestimation of heavy and bulky items, which were the primary source of revenue leakage. Overall, the production pilot achieves a 22% reduction in average per-order revenue leakage while maintaining low abstention rates, confirming that multimodal in-context inference remains reliable under real-world catalog diversity and scale.

Table 3: Pre- and post-pilot Impact

Metric	Pre	Post	Δ
Acc	∼37%	∼**71%**	+34%
UC	∼70%	∼**34%**	-36%
UC (2+)	∼40%	∼**5%**	-35%
RevL	—	—	-22%

5.2 Deployment Insights

- **Correction of High-Cost Errors:** The largest improvements occur in higher weight buckets, where misestimation has the greatest financial impact. The model reduces UC in these segments by 35%, demonstrating effective calibration for heavy products.

- **Operational Robustness:** Despite visual and metadata variability, the system maintains stable accuracy and low AR, proving the LVLM generalizes effectively to unseen listings.

- **Downstream Impact:** Integrating these accurate predictions into product ranking and exposure algorithms will ensure that shipping costs are reflected within the platform's competitive landscape.

6 Conclusion & Future Work

We presented a scalable approach for grounded multimodal attribute inference using LVLMs and ICL, specifically for large-scale e-commerce weight estimation. By conditioning predictions on distribution-calibrated exemplars, our method enables models to infer implicit physical properties from visual and textual cues. Offline evaluations and live production deployment show it consistently outperforms strong multimodal baselines, reducing RevL while maintaining efficiency. Because the approach relies on inference-time conditioning rather than task-specific fine-tuning, it supports rapid onboarding of new categories without retraining. Our results indicate that multimodal ICL is a practical, cost-effective alternative to specialized hardware or manual verification. Future work will explore adaptive exemplar selection, improved calibration, and expansion to attributes like dimensions using more efficient multimodal backbones.

References

Bain & Company. 2025. Ai's trillion-dollar opportunity. https://www.bain.com/insights/ais-trillion-dollar-opportunity-tech-report-2024/.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Eshopbox. 2025. Minimizing shipping overcharges: A guide to tackling weight discrepancies. https://www.eshopbox.com/blog/minimizing-shipping-overcharges.

Jiaying Gong and Hoda Eldardiry. 2024. Multi-label zero-shot product attribute-value extraction. In *Proceedings of the ACM Web Conference 2024*, pages 2259–2270.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Rattapol Kasemrat and Tanpat Kraiwanit. 2024. Benchmarking machine learning models for predictive analytics in e-commerce. *Available at SSRN 4832967*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pages 100–114.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. *arXiv preprint arXiv:2309.10954*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Redseer. 2024. India's festive e-commerce market 2024: Unveiling growth trends and emerging opportunities. https://redseer.com/reports/indias-festive-e-commerce-market-2024/.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 2655–2671.

Shiprocket. 2025. How to reduce weight discrepancies – hacks for 2026. https://www.shiprocket.in/blog/reduce-weight-discrepancies/.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. Flava: A foundational language and vision alignment model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629.

Ankush Verma, Chetan Nagar, Sharda Haryani, and Sumit Jain. 2025. Prediction of e-commerce shoppers' purchasing intention using knn algorithm. In *International Conference on Recent Advancements and Modernisations in Sustainable Intelligent Technologies and Applications (RAMSITA 2025)*, pages 63–72. Atlantis Press.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. Learning to extract attribute value from product via question answering: A multi-task approach. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 47–55.

Huimin Xu, Wenting Wang, Xinnian Mao, Xinyu Jiang, and Man Lan. 2019. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1049–1058.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. *arXiv preprint arXiv:2009.07162*.

Gita Ziabari. 2025. Machine learning to detect abnormal delivery performance in supply chain operations.

Henry Peng Zou, Gavin Heqing Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024. Eiven: Efficient implicit attribute value extraction using multimodal llm. *arXiv preprint arXiv:2404.08886*.

A Dataset Statistics

Refer Table 4 for Dataset statistics.

Table 4: Dataset statistics

SSCat	12 Months	6 Months
Dumbbells	10019	6152
Mops With Bucket	95240	60721
Spice Racks	119634	237050
Collapsible Wardrobe	11585	33458
Collapsible Shoe Racks	138810	161673
Storage Drawer Units	867332	546695
Study Table	251523	1631754
Kitchen Storage	212436	320226
Racks & Holders	77119	248846
Detergents	130723	204324
Plant Containers	121773	170408
Drying Racks	4399	21282
Compost & Fertilizers	25523	24384
Jars & Containers	501640	390588

B Prompt Architecture

For each query product, we construct a single-turn multimodal prompt compatible with Gemini 2.5 Flash. The prompt consists of (i) a task instruction, (ii) a set of labeled exemplars retrieved from the same category, and (iii) the query image–text pair.

Instruction. The system instruction specifies the prediction task and constrains the output space to a predefined set of discrete weight buckets:

> *"You are a model that predicts the most appropriate weight bucket (in grams) for a product. Only select a value from the predefined set of buckets: [500, 1000, ..., 25000]."*

Exemplar Structure. Each exemplar $\mathcal{E}_C$ is interleaved as a tuple (I_i, T_i, y_i) to provide explicit demonstrations of the mapping from multimodal evidence to weight buckets:

- **Textual metadata:** product titles and descriptions conveying material and dimensional cues,
- **Visual inputs:** product images providing size, density, and texture information,
- **Labels:** verified weight buckets serving as calibration references.

This formulation enables the LVLM to perform comparative reasoning by aligning the query against category-specific references within the same context.

Inference Configuration. To ensure consistent and reproducible outputs across large-scale evaluation, we adopt the following settings:

- **Structured outputs:** responses are generated in JSON format (e.g., `application/json`) to facilitate deterministic parsing of the predicted `weight_bucket`.
- **Reasoning budget:** a modest internal reasoning budget is enabled to allow stable inference for multimodal comparisons.

- **Safety configuration:** standard safety thresholds are used with adjustments to avoid false positives on benign e-commerce product images (e.g., household tools), while maintaining policy compliance.

Refer Figure 5 for an illustration of the final prompt structure.

C Additional Qualitative Analysis

We provide qualitative case studies to better understand the behavioral differences between embedding-based retrieval and multimodal in-context inference. The examples below use real product images and metadata from the catalog and illustrate representative success and failure modes across several high-variance Sub-Sub-Categories (SSCats). Table 5 summarizes prediction outcomes, with corresponding visual references shown in Figure 6.

Material and Density Cues. In the Dumbbells category (Figure 6g, 6h), the KNN baseline often relies on global visual similarity and struggles to differentiate products with comparable size but different material density. In contrast, the LVLM-based approach leverages texture and structural cues to better estimate mass, resulting in predictions closer to the ground-truth bucket.

Structural Reasoning and Volume. For Storage Drawer Units (Figure 6d, 6e), embedding-based retrieval frequently matches heavy multi-layer units to visually similar but smaller variants. The multimodal model instead incorporates volumetric and structural cues, producing more accurate weight estimates for taller, denser configurations.

Integration of Textual and Visual Signals (OCR). In Compost & Fertilizers (Figure 6l), the baseline matches items primarily based on packaging appearance. The LVLM additionally leverages textual cues visible in the image (e.g., "20KG"), enabling more reliable bucket assignment. This example highlights the benefit of combining visual grounding with embedded text recognition.

Industrial vs. Household Variants. In the Mops with Bucket category (Figure 6a, 6b), the multimodal model distinguishes industrial-grade components such as metal wringers and thicker structural elements from lighter household variants. This results in improved calibration for heavier products that are systematically underestimated by embedding-only baselines.

Table 5: Category-wise bucket prediction comparison between KNN and our method

SSCat	Image	Description	Actual (kg)	KNN bucket (kg)	Ours bucket (kg)
Mops With Bucket	6a	CRIZAR 360° Spin Floor Cleaning Easy Advance Tech Bucket Mop WITH WIPER & Rotating Steel Pole Head, Mop Set Blue Colour (With 1 Refill)	7.5	4.5–5	7.5–8
	6b	Zelenor Easy Spin Floor Cleaning Bucket Mop with 2 Super Microfiber Absorber, Tile Scrub, Liquid Dispenser and 4 In 1 Foot Pedicure Brush includes Pumice Stone, Scrubber & File for Home	8	13–13.5	8–8.5
	6c	The Flying Tree Bucket Quick Spin Mop with 2 Microfiber Wet Dry Mophead Floor Cleaning pocha Extendable Handle Removable Wringer 360° Floor Cleaner Mopping Set, Mop Bucket, Mopping Bucket, Mopping Machine, Mopping Stick Mopping Set	6	14–14.5	6.5–7
Storage Drawer Units	6d	ELIGHTWAY MART Easy Storage Modyular Supar Drawer(7XL-SKUBLUE)	12	8–8.5	11–11.5
	6e	Plastic Drawers Storage - Versatile Storage Drawer Organizer Your Essentials - Ideal Drawers for Storage in Home, Office, or Classroom Brown | 7 Layer (7 Layer Brown)	14	10–10.5	13.5–14
	6f	5 XI Plastic Modular Drawer System For Home, Office, Hospital, Parlor, School, Doctors, Home And Kids, Product Dimension When Assemebled (36Cmx39Cmx98Cm)(5XI-Classic), Multi	6	9–9.5	6–6.5
Dumbbells	6g	PVC Dumbbells Set For Home Gym, Exercise & Fitness 3Kg Pair Hexa	6.5	8.5–9	6.5–7
	6h	RA HEXA Dumball 4kg X2 pcs = 8 kg set For Home Workout and Fittess	8.5	4–4.5	8.5–9
	6i	20 kg of PVC weight (3 kg x 4 = 12 kg, 2 kg x 4 = 8 kg)with gloves and Hand gripper And skipping rope 2 x 14 inch dumbbell rods with nuts	21	13–13.5	21.5–22
Compost & Fertilizers	6j	Cocopeat Block for Home Garden Plants -9.9 kg block (Set of 4.9 KG Each, 10.5 x 15 cm),Soil Manure Potting Mixture for Home Garden, soil,manure-Ecocoir	9	5–5.5	8.5–9
	6k	Combo Pack Bio DAP(5Kg)+Bio Potash(5Kg)	10.5	14–14.5	10–10.5
	6l	vermicompost | cow dung manure | plant booster | plant food	20.5	11–11.5	19.5–20
kitchen Storage	6m	Kitchen & Home Storage for Onion, Potato, Aloo, Pyaaz, Fruit, Vegetable, Sabji & Fish, Fruits & Veggies, Ideal And Clothes Storage Underwear, Socks, Scarf, t-Shirt, Ideal for Home Use, Multipurpose Use Best Trolley, Best In Household Products,Useful in kitchen,Storage Trolley,Kitchen Trolley- WHITE	6	9–9.5	6.5–7
	6n	rack 4 tier	7	4.5–5	7–7.5
	6o	3 Layer Kitchen Trolley Storage Rack Square Design Fruits & Vegetable Basket metal Kitchen Trolley (Pre-assembled)	11.5	6–6.5	11.5–12

Final Multimodal Prompt Architecture

System Instruction:
`You are a machine learning model that predicts the most appropriate weight bucket (in grams) for a product. The prediction is based on an image and a text description of the product, using a few examples from the same category as reference. Only select a weight from the predefined set of buckets: [500, 1000, ..., 25000].`

User Content:
You are given example products with their descriptions, images, and weight buckets. Use these to predict the weight bucket of the target product.

Examples:

- **Product 1 Description:** CRIZAR 360° Spin Floor Cleaning Easy Advance Tech Bucket Mop WITH WIPER & Rotating Steel Pole Head, Mop Set Blue Colour (With 1 Refill)
  Product 1 Image: 
  Product 1 Weight bucket: 7500 grams
- .....
- .....
- .....
- **Product 50 Description:** Zelenor Easy Spin Floor Cleaning Bucket Mop with 2 Super Microfiber Absorber, Tile Scrub, Liquid Dispenser and 4 In 1 Foot Pedicure Brush includes Pumice Stone, Scrubber & File for Home
  Product 50 Image: 
  Product 50 Weight bucket: 8000 grams

Target Product:
Description: The Flying Tree Bucket Quick Spin Mop with 2 Microfiber Wet Dry Mophead Floor Cleaning pocha Extendable Handle Removable Wringer 360° Floor Cleaner Mopping Set, Mop Bucket, Mopping Bucket, Mopping Machine, Mopping Stick Mopping Set
Product Image: 
Query: What is the predicted weight bucket (in grams) of the above target product?

Generation Config (JSON Schema Enforcement):

```
{
  "type": "object",
  "properties": {
    "weight_bucket": {
      "type": "integer",
      "enum": [500, 1000, 1500, ..., 25000]
    }
  },
  "required": ["weight_bucket"]
}
```

Figure 5: Final Prompt Architecture

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

(k)

(l)

(m)

(n)

(o)

Figure 6: Qualitative Analysis.