

A Appendix

A.1 Dataset Characterization	17
A.1.1 Font Size of Product Description Text in Images	17
A.2 Data Structure in Repository	18
A.3 Image and Text Classification	19
A.4 Evaluation of Embeddings in the Visual RAG Pipeline	19
A.5 Evaluation Results on VLMs	20
A.6 Evaluation Results on Visual RAG Pipeline	20
A.7 Further Error Analysis and Limitations of VLM Based and Visual RAG Based Approaches	20
A.8 Fine-Tuning Based Approaches	21
A.9 Extra file	22

A.1 Dataset Characterization

Figure 7 shows the distribution of the number of images per label in the *mSOP-765k* dataset. The ratio of frequency and accuracy of the GTINs count per advertisement is illustrated in Figure 8. The accuracy score is based on the evaluation of the VLM-based approach of the model gemini-2.5-flash. The OCR-extracted text of an image from the dataset and the image itself are presented in Figure 9. Appendix A.1.1 discusses the font size of the text in the images.

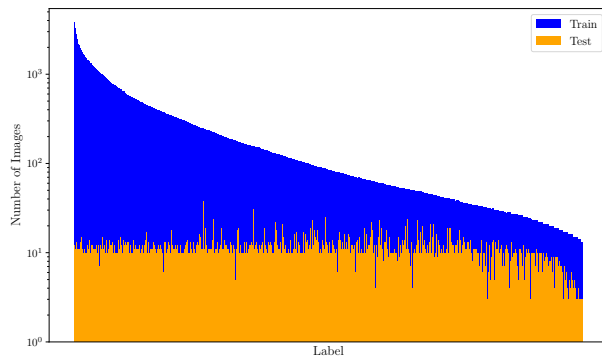


Figure 7: Histogram illustrating the image distribution in the train and test split of the *mSOP-765k* dataset.

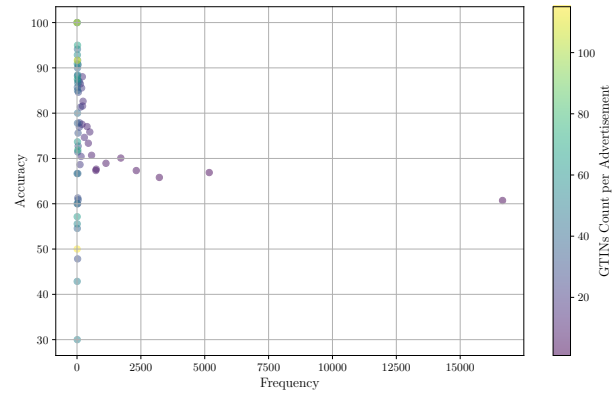


Figure 8: Illustration of the frequency and accuracy ratio for the GTINs count per advertisement (VLM-based with gemini-2.5-flash).

A.1.1 Font Size of Product Description Text in Images

The images included in the dataset were cropped from high-resolution leaflets (Lamm & Keuper, a). From these, the samples presented in Figure 5 are selected as representative examples for analysis. Our focus is directed towards the text with the smallest font size, which consistently corresponds to the weight information across all images. Specifically, the font sizes observed are 18 pixels (Figure 5a), 23 pixels (Figure 5b), and 28 pixels (Figure 5c). Taking into account these font dimensions alongside the text classification accuracy of 78.5%, as detailed in Appendix A.3, the font size can be regarded as adequate.



(a) Image from the dataset.

Zotto ZOTT Aktion Zottarella* Mozzarella Classic
oderJalapeno, je200-g-Beutel, 0.89 Abtropfgewicht= 125g
100g=0.71

(b) OCR-extracted text of the promotion image shown in Figure 9a.

Figure 9: Illustration of an image from the *mSOP-765k* dataset (9a) and its OCR-extracted text (9b).

A.2 Data Structure in Repository

The *mSOP-765k* benchmark provides the product and promotion data for each advertisement image. For each split, the data are made available in the repository in the `.parquet` file format. The file size of `train.parquet` is approximately 5.6MB, while the file `test.parquet` is about 646kB. The column names in these files are as follows: `label`, `filename`, `brand`, `price`, `regular_price`, `relative_discount`, `absolute_discount`, `product_category`, `GTINs`, `product_weight`, and `different_types`. The GT values for the target *product category* may consist of a list of values. The categories are not printed on the advertisement images; rather, they are predefined by an internal system. For the promotion image shown in Figure 2a, the product category is defined as: `Suppentopf`, `Suppen`, `Suppe` (engl. stockpot, soups, soup). All product and promotion data to the Figure 2a are illustrated in Table 2b. Similarly, the target *GTINs* may comprise a list of values and are never printed on the advertisement images. The target variable *different types* is set to `yes` if there is evidence of different types in the promotion image; otherwise, this field is left empty, i.e. `NaN`. Table 9 shows, for each target variable, whether a prediction is required and specifies the data type used.

The repository contains the folders `rpp-765k_256` and `rpp-765k_512`, which provide image datasets with the longer image edge resized to 256 and 512 pixels, respectively. The `rpp-765k_256` image dataset has a size of about 7.9GB, while the `rpp-765k_512` dataset is approximately 59GB. Each version of the image dataset is divided into `train` and `test` split, respectively. Each split folder contains compressed files in the `.tar.gz` format, with each archive holding advertisement images corresponding to each label.

The files `OCR_train.parquet` (approx. 2.64MB) and `OCR_test.parquet` (approx. 37.4MB) contain the OCR-extracted text for each advertisement image.

Table 9: Overview of each variable data type and indication of whether prediction is required.

Target	Data Type	Required
brand	string	✓
product category	list of strings	✓
GTINs	list of strings	✓
weight number	float	✓
weight unit	enum	✓
different types	enum	✓
price	float	✓
regular price	float	✗
relative discount	float	✗
absolute discount	float	✗

Table 10: Overview of the numbers of images, image embeddings, and text embeddings for the label with a representative example illustrated in Figure 9a.

	images	image embeddings	text embeddings
train split	67	90	66
test split	10	12	9

A.3 Image and Text Classification

These evaluations primarily focus on the images and OCR-extracted texts of the *mSOP-765k* dataset. The image-only and text-only classifications were performed on an NVIDIA GeForce RTX 4090 GPU in each case. In addition, the training settings as well as the hyperparameter used for the classifications followed those used in the baseline experiments by Lamm & Keuper (b).

Image Classification: For the baseline image-only classification, the ResNet50 model (He et al.) is used. The training process took approximately 36 hours and the image classification has achieved an accuracy of 83.0% and an F1 score of 80.3%.

Text Classification: The texts printed on the images were used. Text extraction was performed using the PaddleOCR (PaddlePaddle; Du et al.) tool. The BERT model (Devlin et al.) is evaluated. Training the text classifier took approximately 21 hours. The accuracy and F1 score achieved have been 78.5% and 72.8%, respectively.

Error Analysis of Image and Text Classification: The total number of misclassifications resulting from image classification has been 2,992. An aggregate of 4,724 misclassifications have been observed from the text classification. The number of erroneous classifications common to both image and text classification has been 1,799. The majority of misclassifications in both image and text classification have occurred when images are assigned labels that differ only in weight content or weight specification.

A.4 Evaluation of Embeddings in the Visual RAG Pipeline

Figure 10 and Figure 11 illustrate the distributions of cosine distances for image and text embeddings across different observation sets, i.e., specifically distinguishing between intra-label distances (Figure 10) and distances measured between training and test splits (Figure 11). For calculations, we focus on a single label with a representative example illustrated in Figure 9a. The considered embeddings are stored in the Vector Store of the Visual RAG Pipeline. Table 10 presents the overview of the numbers of images, image embeddings, and text embeddings for the label considered. Figure 11 shows that the distances between the text embeddings of the train and the test splits are significantly higher compared to the distances within the same label. Moreover, the distances between image embeddings are distributed more evenly when comparing intra-label and cross-split distributions.

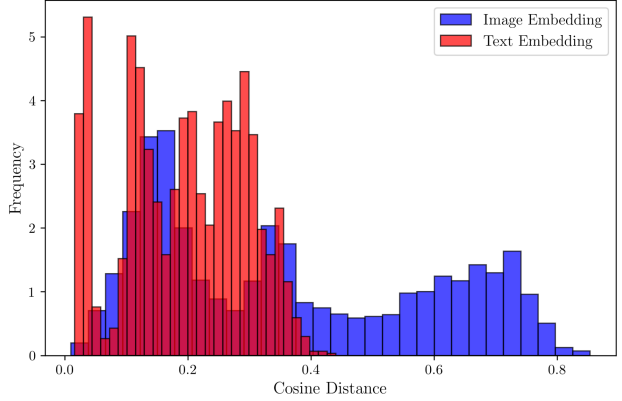


Figure 10: An illustration of the distribution of intra-label cosine distances calculated separately among image embeddings and among text embeddings for a single label, with all embeddings stored in the Vector Store of the Visual RAG Pipeline.

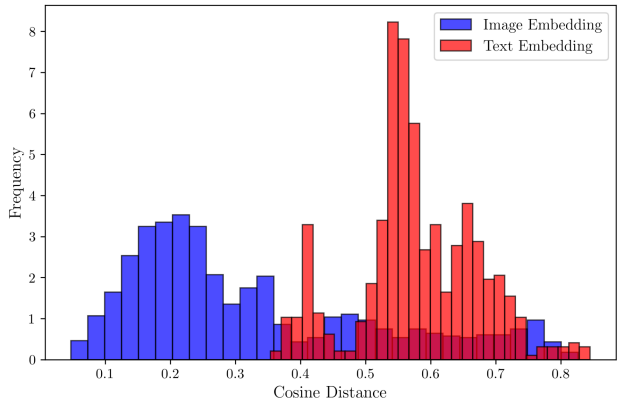


Figure 11: An illustration of the distribution of cosine distances calculated separately among image embeddings and among text embeddings for a single label, with distances calculated between embeddings from the training and test sets; all embeddings are stored in the Vector Store of the Visual RAG Pipeline.

A.5 Evaluation Results on VLMs

Figure 13 illustrates the difference between the number of predictions and the number of GT values per each VLM. Further results on the evaluation metrics for the single target variables and for the structured output are presented in Table 4 and Table 14, respectively. The detailed plot of the prediction errors for the target *price* is shown in Figure 15. The evaluation of the elapsed time as well as the total costs is summarized in Table 15.

For some metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), evaluations can only be performed if the prediction and GT of the variable are defined. The percentage of analyzable data is shown in Table 16. The VLM gpt-4o-mini-2024-07-18 has delivered the best result in most targets. For the target *product weight* all VLMs have had the same percentage. This results from the configuration that the target *product weight* is required, i.e. there must be a prediction. The low percentage for the targets *regular price*, *relative discount*, and *absolute discount* results from the low number of GT values in the test split. Such variables are not provided for every advertisement image.

Table 17 presents the metrics MSE and RMSE of the promotion data for each VLM. The lowest values for both metrics have been shown by the VLMs gpt-4o-mini-2024-07-18 and gemini-2.0-flash. The VLM llama-3.1-8b has indicated huge values for both metrics and for all target variables.

The VLMs gemini-2.5-flash and gpt-5-mini-2025-08-07 provide reasoning tokens in addition to the input and output tokens. Table 11 shows the mean, minimum, and maximum number of reasoning tokens for the VLMs gemini-2.5-flash and gpt-5-mini-2025-08-07.

Table 11: VLM-based – Illustration of the mean, minimum, and maximum number of reasoning tokens for the VLMs gemini-2.5-flash and gpt-5-mini-2025-08-07.

	gemini-2.5-flash	gpt-5-mini-2025-08-07
Mean Tokens	1,485.58	830.65
Minimum Tokens	136	192
Maximum Tokens	62,916	2,688

A.6 Evaluation Results on Visual RAG Pipeline

Table 18 shows the percentage of analyzable data, i.e. the GT and prediction values are defined. The VLM gpt-4o-mini-2024-07-18 has almost achieved the best value in all the targets except the target *GTINs*. For this target the VLM mistral-small3.1-24b has attained a percentage score of 98.14%.

Table 19 illustrates the metrics MSE and RMSE for each VLM used in the Visual RAG Pipeline. The VLM gpt-4o-mini-2024-07-18 has gained the best MSE and RMSE metrics for the targets *price*, *regular price*, and *relative discount*. For the target *absolute discount*, the VLM mistral-small3.1-24b has achieved the lowest MSE value of 12.06 and RMSE value of 3.47.

Figure 17 displays the difference between the number of predictions and the number of GT values for the targets and each VLM used in the Visual RAG Pipeline. For all targets except the targets *regular price*, *relative discount*, and *absolute discount*, the number of missed predictions has been the same as the number of failed requests in step *Completion* of the Visual RAG Pipeline.

A.7 Further Error Analysis and Limitations of VLM Based and Visual RAG Based Approaches

Tables 22 to 24 show the GT values and the SOPs from the approach combinations VLM/gemini-2.0-flash and Visual RAG Pipeline/gpt-4o-mini-2024-07-18 for the advertisement images displayed in Figure 5. Table 25 shows the same data for Figure 6. Figures 18 to 20 present examples that illustrate the limitations of the investigated approaches.

Figure 19b shows the price predictions for each model used the VLM-based approach. The response of the VLM llama3.2-vision-11b misses a value for the target *price*, i.e. the value NaN is returned. Hence, no evaluation can be calculated for this target. For the evaluation metric \bigcup_{targets} considering the target *price*, this image is skipped. For the evaluation metric \bigcup_{test} , the missing prediction is considered as false prediction.

Table 23 presents the GT values and the predictions of Figure 5b of a VLM based and a Visual RAG based approach. The GT value of the target *GTINs* consists of a list of values. The VLM based approach returns no prediction for this target because the information is not printed in the advertisement image. The Visual RAG based approach predicts the correct list of GTINs. A prediction for the target *GTINs*, given as a list, is considered correct only if all entries in the predicted list match the GT values exactly and appear in the same order.

A.8 Fine-Tuning Based Approaches

In this section we evaluate alternative approaches to RAG, using fine-tuning of VLMs. Due to computational constraints and training set size constraints issued by OpenAI (maximum training set size is limited to 512MB for finetuning), we created a random subset of the training dataset, which is used to fine-tune VLMs. This random training dataset consists of 10,250 images that belong to 2,010 different labels. The training dataset includes the structured outputs of all targets, particularly the targets *product category* and *GTINs*. Moreover, a validation subset of 2,500 images is provided. The supervised fine-tuning process of OpenAI (OpenAI, e) is applied. It is important to note that some images are not included in the fine-tuning training process, due to OpenAI’s policy regarding image content (OpenAI, f). The policy forbids images showing people or faces. Figure 12 shows an image example that is excluded. The majority of the default settings for the training process are used, we adapt the hyperparameter number of epochs to 3. The costs of the fine-tuned training process amounts to about USD 340.00. The fine-tuning training process has taken approximately 5.6 hours. For the evaluation, the model GPT-4o in the version dated 08/06/2024 (gpt-4o-2024-08-06) (OpenAI, b) and the fine-tuned model (gpt-4o-2024-08-06_FT) are considered.



Figure 12: Image excluded from model fine-tuning by OpenAI’s image content policy.

Further, a local fine-tuning is performed with the open-source VLM Gemma 3 4B (gemma-3-4b) (Google, b). The resulted model is called: `gemma-3-4b-local_FT`. The HuggingFace Transformers (HuggingFace) and QLoRA (Dettmers et al.) are used to fine-tune the base model. As training data, the same random subset of 10,250 images is used. Also, the same validation subset is applied. The most settings of the coding example (Google, a) are adopted. The following settings are changed: number training epochs = 3, training / evaluation batch size per device = 2, and the use of an evaluation dataset for training. The fine-tuning process is executed on 4 NVIDIA A100-SXM4-40GB. The duration of the fine-tuning training process with the random subset for the VLM gemma-3-4b has taken approximately 11.4 hours. Therefore, it would take around 34 days to fine-tune the model with the entire training dataset.

Table 13 presents the scores of the evaluation metric for the targets *product category* and *GTINs*. The fine-tuned VLM gpt-4o-2024-08-06_FT has delivered only 6.84% right predictions for the target *product category*. For the target *GTINs*, no right predictions have been generated. However for this target, the fine-tuned VLM gemma-3-4b-local_FT has produced 5.60% right predictions. For the target *product category*, the model has delivered 54.41% right predictions. Compared to Table 7, these values are significantly lower than the evaluation using the Visual RAG based approach. The scores of the evaluation metrics for each single target variable and VLM with / without fine-tuning are presented in Table 26. Moreover, Table 27 and Table 28 show the evaluation metrics \bigcup_{targets} and \bigcup_{test} , respectively. The fine-tuned version gpt-4o-2024-08-06_FT has produced more false predictions in both evaluation metrics compared to the VLM gpt-4o-2024-08-06. The model gemma-3-4b-local_FT has demonstrated improved evaluation metrics compared to its base model without fine-tuning. However, the evaluation metrics have remained at 0% for both versions when the target *relative discount* was included. Table 29 shows that each fine-tuned VLM has required more time and has incurred higher costs compared to its base model.

Table 12 presents the mean token accuracy of the training dataset for the VLMs with and without fine-tuning. The accuracy of the fine-tuned models has increased compared to the models without fine-tuning.

Table 12: VLM-based With/Without Fine-Tuning – Illustration of the mean token accuracy on the training dataset.

	gpt-4o- 2024-08-06	gpt-4o- 2024-08-06_FT	gemma-3-4b	gemma-3-4b- local_FT
Mean Token Accuracy	70.2%	<u>94.7%</u>	69.2%	98.6%

Table 13: VLM-based With Fine-Tuning – Scores of the evaluation metrics for the targets *product category* and *GTINs* and the fine-tuned models. Percentages are calculated based on comparable GT and predicted values.

	gpt-4o- 2024-08-06_FT	gemma-3-4b- local_FT
product category	6.84%	54.41%
GTINs	0%	5.60%

Table 14: VLM-based – Results of the evaluation metric \bigcup_{targets} for the structured output. The scores are calculated by systematically incorporating additional targets, based on the set of comparable GT values and predictions.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	llava-34b	qwen2.5 vl-32b	mistral- small3.1-24b	llama3.2- vision-11b	gemini- 2.5-flash	gpt-5-mini- 2025-08-07
brand	95.55%	<u>96.61%</u>	70.79%	93.37%	94.12%	90.77%	96.73%	95.82%
\bigcup_{targets} product weight	68.35%	78.74%	5.78%	16.12%	16.39%	18.46%	70.36%	<u>77.39%</u>
\bigcup_{targets} different types	65.18%	76.89%	8.30%	11.51%	12.43%	16.54%	68.57%	<u>76.87%</u>
\bigcup_{targets} price	63.02%	<u>74.59%</u>	4.45%	10.92%	11.84%	15.63%	66.58%	74.79%
\bigcup_{targets} regular price	55.33%	<u>68.83%</u>	1.50%	4.79%	8.17%	2.20%	64.61%	69.95%
\bigcup_{targets} relative discount	58.01%	<u>67.89%</u>	0.64%	8.00%	7.32%	0%	64.66%	70.25%
\bigcup_{targets} absolute discount	0%	0%	0%	0%	0%	0%	0%	0%

However, the evaluations on the test dataset show a decline in the evaluation metrics (see: Tables 26 to 28). Therefore, the fine-tuned models are overfitted to the training dataset.

A.9 Extra file

The dataset is documented in the machine-readable format: Croissant (MLCommons). The appropriate file, `mSOP-765k_croissant_metadata.json`, is provided in the supplemental material.

²Usage of local hardware; costs are estimated on the assumption of the usage of a Linux VM D11 v2 with 2 vCPUs in Azure: costs are USD 135.05/month (Microsoft).

Table 15: VLM-based – Illustration of the average elapsed time and costs per request as well as the total elapsed time and costs across all requests for each VLM.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	llava-34b	qwen2.5 vl-32b	mistral- small3.1-24b	llama3.2- vision-11b	gemini- 2.5-flash	gpt-5-mini- 2025-08-07
avg. elapsed time per req.[s]	11.60	2.15	5.03	<u>4.56</u>	26.55	25.52	6.36	16.94
total elapsed time (all req.)[h]	117.9 (= 4.9d)	21.7	51.1	<u>46.2</u>	260.4 (= 10.6d)	242.5 (=10.1d)	64.6	172.0 (=7.2d)
avg. total costs per req.[USD]	0.0014	0.0002	<u>0.0003</u> ²	0.0002 ²	0.0014 ²	0.0013 ²	0.0021	0.0020
total costs (all req.)[USD]	50.47	7.36	9.59 ²	<u>8.67</u> ²	48.85 ²	45.49 ²	75.86	73.23

Table 16: VLM-based – Percentage of images from the test split for which GT and prediction can be analyzed, i.e., GT and prediction values have non-null entries.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	llava-34b	qwen2.5 vl-32b	mistral- small3.1-24b	llama3.2- vision-11b
brand	99.98%	<u>99.55%</u>	99.12%	99.54%	96.53%	78.95%
product weight	96.70%	96.70%	96.70%	96.70%	96.70%	96.70%
price	99.90%	<u>99.47%</u>	99.04%	99.46%	96.45%	78.88%
regular price	35.15%	<u>35.01%</u>	25.32%	2.34%	31.99%	0.84%
relative discount	<u>32.04%</u>	<u>31.90%</u>	39.60%	30.29%	29.78%	19.23%
absolute discount	1.82%	<u>1.81%</u>	1.73%	1.80%	1.75%	1.47%

Table 17: VLM-based – Illustration of the metrics MSE and RMSE of the promotion data for each VLM.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	llava-34b	qwen2.5 vl-32b	mistral- small3.1-24b	llama3.2- vision-11b
price	<u>34.30</u>	7.44	67,921.57	3,598.25	546.44	4,387.88
	<u>5.86</u>	2.73	260.62	59.99	23.38	66.24
regular price	1.06	<u>2.38</u>	137,502.48	7.31	16.81	25.71
	1.03	<u>1.54</u>	370.81	2.70	4.10	5.07
relative discount	<u>26.37</u>	15.00	1,209.20	964.97	57.34	341.06
	<u>5.13</u>	3.87	34.77	31.06	7.57	18.47
absolute discount	14.82	<u>14.15</u>	106.96	27.36	17.61	13.65
	3.85	<u>3.76</u>	10.34	5.23	4.20	3.49

Table 18: Visual RAG-based – Percentage of images from the test split for which GT and prediction can be analyzed, i.e., GT and prediction values have non-null entries.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	qwen2.5 vl-32b	mistral- small3.1-24b
brand	98.14%	98.03%	<u>98.06%</u>	98.14%
product category	98.14%	98.03%	<u>98.06%</u>	98.14%
GTINs	<u>98.06%</u>	98.03%	<u>98.06%</u>	98.14%
product weight	96.70%	96.70%	96.70%	96.70%
price	98.06%	97.94%	<u>97.97%</u>	98.06%
regular price	34.46%	<u>34.42%</u>	14.29	17.60%
relative discount	31.46%	<u>31.43%</u>	12.14%	15.23%
absolute discount	1.79%	1.79%	0.28%	<u>0.42%</u>

Table 19: Visual RAG-based – Illustration of the metrics MSE and RMSE per promotion target variable and VLM.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	qwen2.5 vl-32b	mistral- small3.1-24b
price	7.97	<u>12.97</u>	20.09	20.25
	2.82	<u>3.60</u>	4.48	4.50
regular price	7.44	<u>7.80</u>	16.91	15.79
	2.73	<u>2.79</u>	4.11	3.97
relative discount	21.89	<u>56.93</u>	142.57	428.81
	4.68	<u>7.55</u>	11.94	20.71
absolute discount	14.00	<u>13.83</u>	19.39	12.06
	3.74	<u>3.72</u>	4.40	3.47

Table 20: Visual RAG-based – Results of the evaluation metric \bigcup_{targets} for the structured output. The scores are calculated by systematically incorporating additional targets, based on the set of comparable GT values and predictions.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	qwen2.5 vl-32b	mistral- small3.1-24b	gemini- 2.5-flash
brand	<u>98.53%</u>	95.21%	90.04%	90.08%	98.81%
\bigcup_{targets} product weight	<u>86.12%</u>	85.77%	79.61%	79.44%	86.43%
\bigcup_{targets} different types	86.86%	86.09%	81.77%	80.87%	<u>86.46%</u>
\bigcup_{targets} price	83.72%	56.87%	18.80%	18.67%	<u>82.80%</u>
\bigcup_{targets} regular price	<u>71.44%</u>	54.92%	11.56%	11.91%	78.32%
\bigcup_{targets} relative discount	<u>73.97%</u>	54.11%	16.07%	16.00%	80.11%
\bigcup_{targets} product category	66.38%	48.10%	14.88%	15.23%	<u>61.14%</u>
\bigcup_{targets} GTINs	59.17%	42.87%	12.80%	14.24%	<u>54.59%</u>
\bigcup_{targets} absolute discount	0%	0%	0%	0%	0%

Table 21: Visual RAG-based – Illustration of the average elapsed time and costs per request as well as the total elapsed time and costs across all requests for each VLM.

	gpt-4o-mini- 2024-07-18	gemini- 2.0-flash	qwen2.5 vl-32b	mistral- small3.1-24b	gemini- 2.5-flash
avg. elapsed time per req.[s]	16.82	4.22	9.14	<u>7.42</u>	12.21
total elapsed time (all req.)[h]	167.7 (= 7.0d)	42.04	91.04 (= 3.8d)	<u>74.02</u>	121.72 (= 5.1d)
avg. total costs per req.[USD]	0.0145	0.0106	<u>0.0005</u> ²	0.0004 ²	0.036
total costs (all req.)[USD]	519.15	380.65	<u>17.08</u> ²	13.89 ²	1,299.86

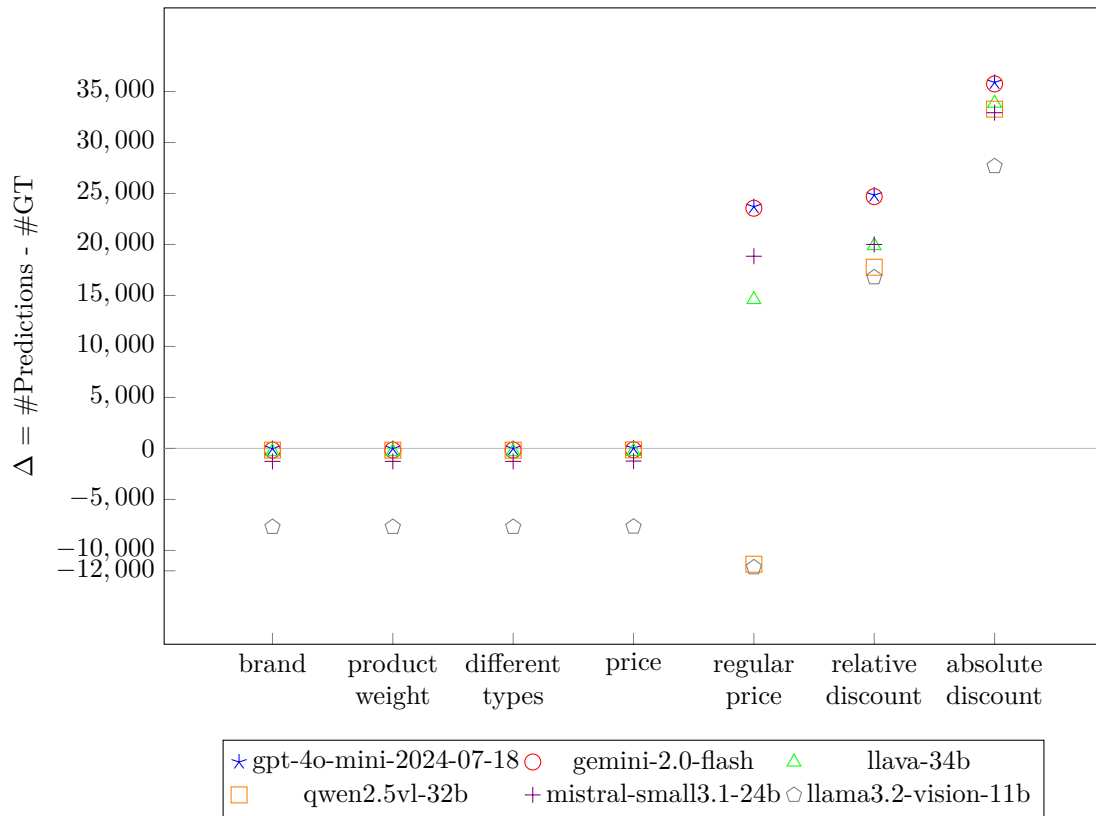


Figure 13: VLM-based – Illustration of the difference between the number of predictions and the number of GT values per target variable and VLM. For the target variables *brand*, *product weight*, *different types*, and *price*, the number of predictions has largely corresponded to the number of GT values for almost all VLMs. For the other target variables, the number of predictions has exceeded the number of GT values by 15,000 to 35,000. The variable *regular price* has been an exception.

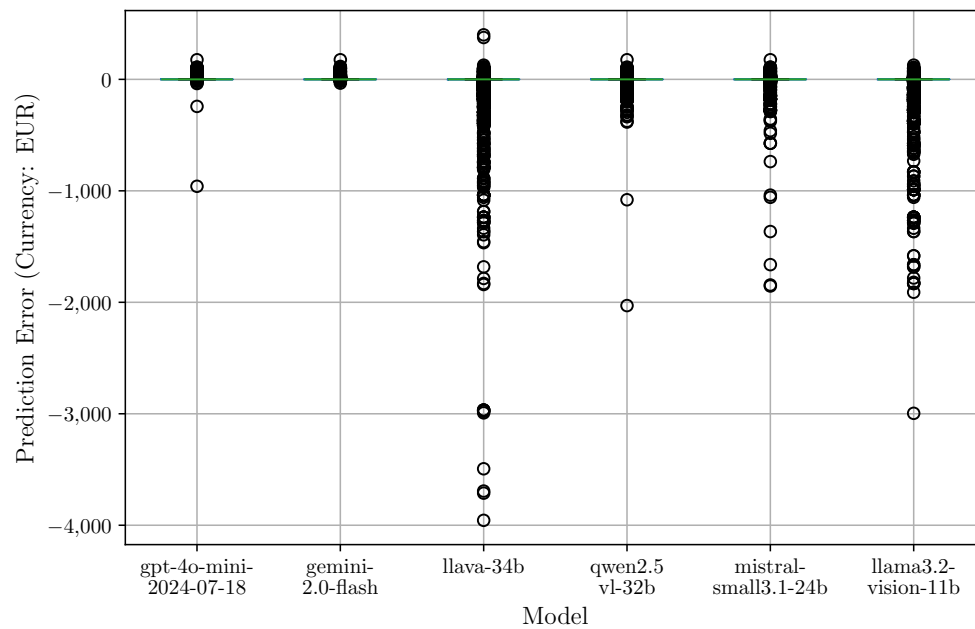


Figure 14: VLM-based – Illustration of the prediction errors for the target *price* per each VLM used. The outliers for the models llava-34b and qwen2.5vl-32b have been truncated, therefore the value range only extends down to -4,000. The complete boxplot visualization is presented in Figure 15.

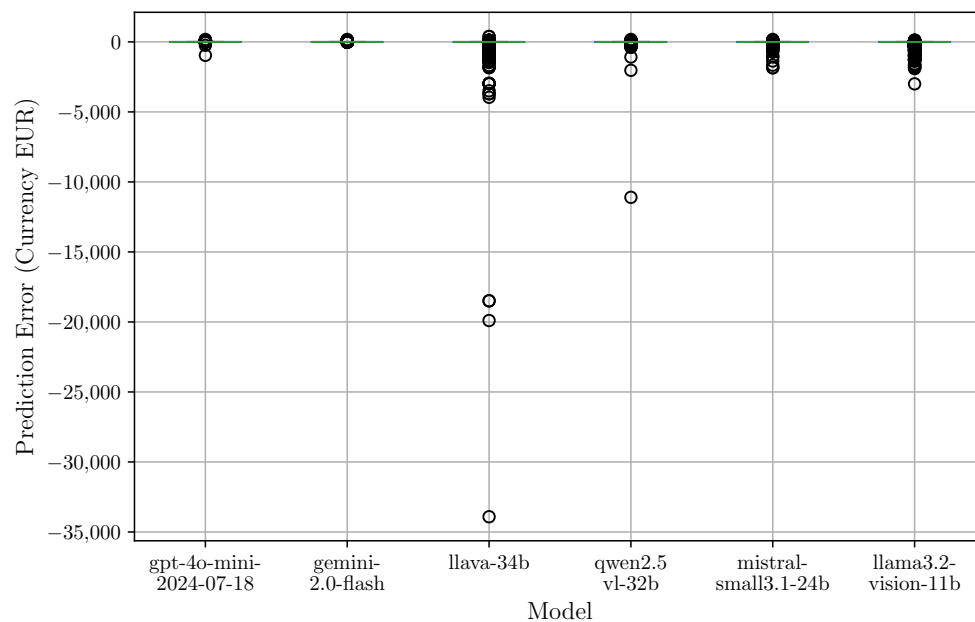


Figure 15: VLM-based – Illustration of the prediction errors for the target *price* per each VLM used.

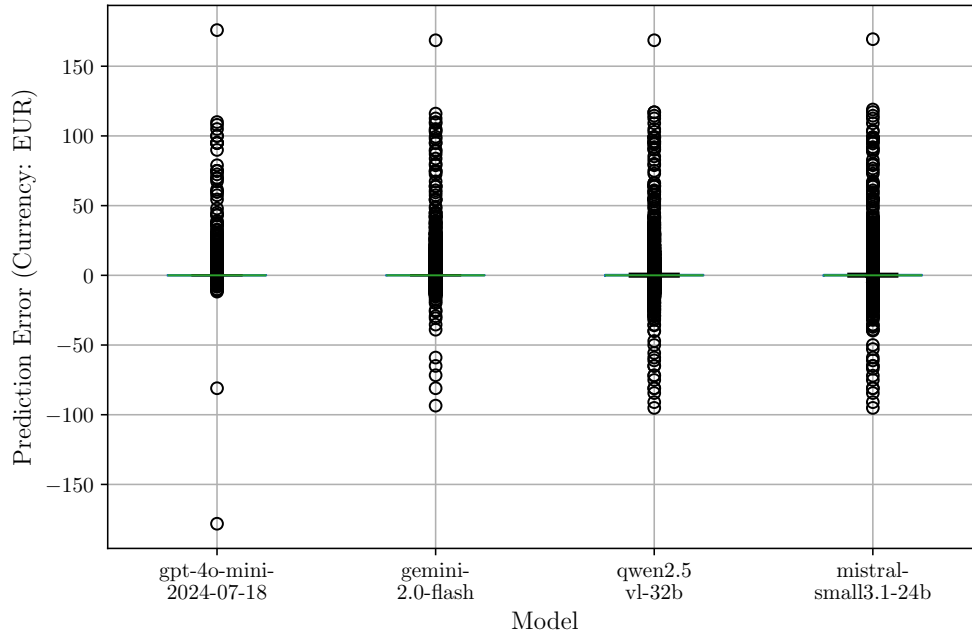


Figure 16: Visual RAG-based – Illustration of the prediction errors for the target price per VLM used in the Visual RAG Pipeline.

Table 22: GT values and SOPs from the methods VLM/gemini-2.0-flash and Visual RAG Pipeline/gpt-4o-mini-2024-07-18 for the promotion image shown in Figure 5a. Incorrect predictions are underlined.

Targets	GT	VLM/gemini-2.0-flash	Visual RAG Pipeline/ gpt-4o-mini-2024-07-18
brand	L’Oreal, Men Expert	L’ORÉAL	L’Oreal
product category	Body Spray, Körperspray, Deo-Spray, Deospray, De- odorant	-	Body Spray, Körperspray, Deo-Spray, Deospray, De- odorant
GTINs	[03600521849552, 03600521849569, 03600521984123]	-	[03600521849552, 03600521849569, 03600521984123]
weight number	150.0	150.0	150.0
weight unit	Milliliter	Milliliter	Milliliter
different types	yes	yes	yes
price	1.69	1.69	1.69
regular price	1.95	1.95	<u>1.99</u>
relative discount	13	13	13
absolute discount	NaN	<u>0.26</u>	<u>0.3</u>

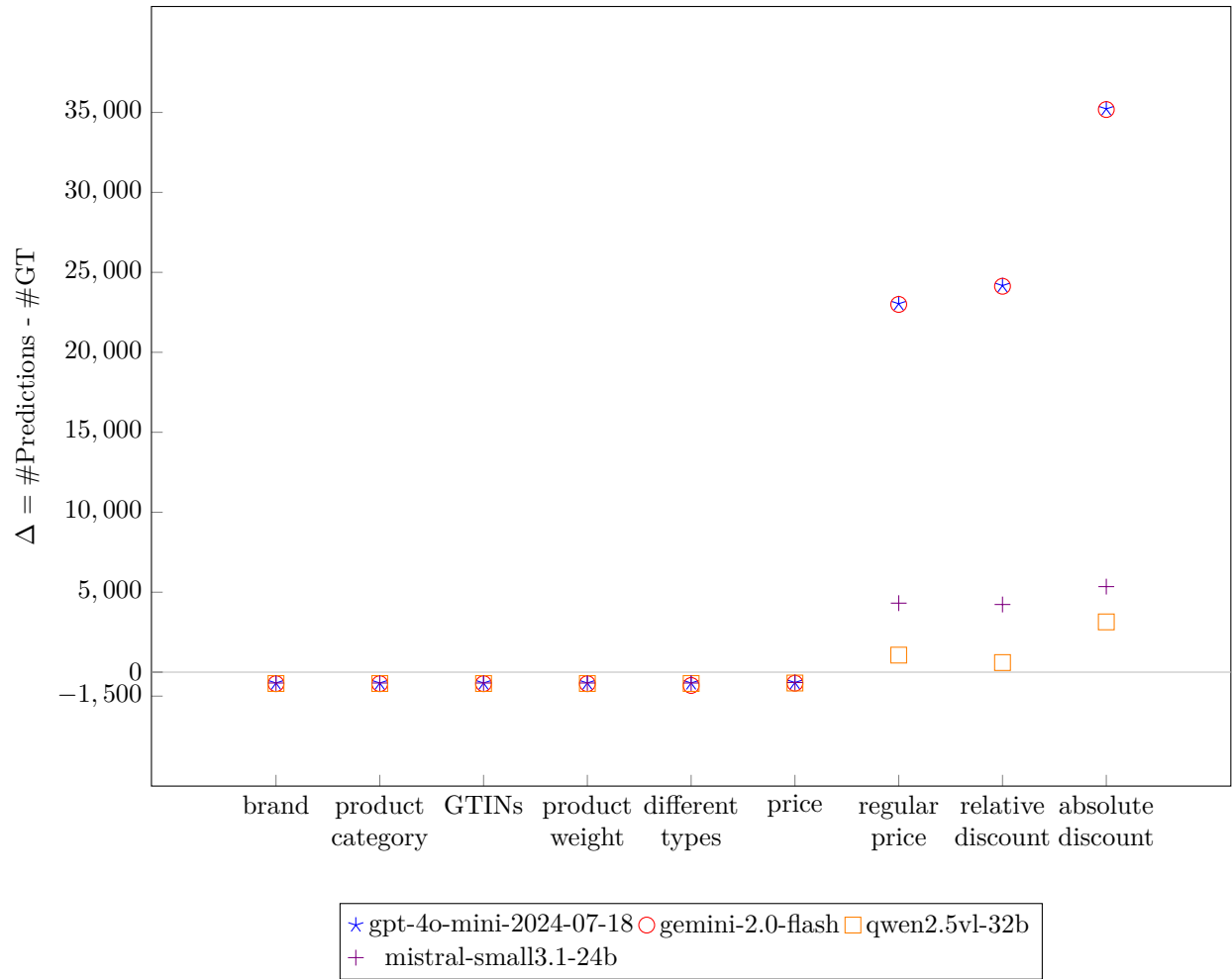


Figure 17: Visual RAG-based – Illustration of the difference between the number of predictions and the number of GT values per target variable and VLM used in the Visual RAG Pipeline.

Table 23: GT values and SOPs from the methods VLM/gemini-2.0-flash and Visual RAG Pipeline/gpt-4o-mini-2024-07-18 for the promotion image shown in Figure 5b. Incorrect predictions are underlined.

Targets	GT	VLM/gemini-2.0-flash	Visual RAG Pipeline/ gpt-4o-mini-2024-07-18
brand	Mentos	Menots	Mentos
product category	Bubble Gum, Kaugummi, Gum	-	Bubble Gum, Kaugummi, Gum
GTINs	[00000080731566, 00000080781073, 00000080906285, 00000080932673]	-	[00000080731566, 00000080781073, 00000080906285, 00000080932673]
weight number	35.0	<u>70.0</u>	35.0
weight unit	Stück	<u>Gramm</u>	Stück
different types	yes	yes	yes
price	1.79	1.79	1.79
regular price	2.49	2.49	2.49
relative discount	28	28	28
absolute discount	NaN	<u>0.7</u>	<u>0.7</u>

Table 24: GT values and SOPs from the methods VLM/gemini-2.0-flash and Visual RAG Pipeline/gpt-4o-mini-2024-07-18 for the promotion image shown in Figure 5c. Incorrect predictions are underlined.

Targets	GT	VLM/gemini-2.0-flash	Visual RAG Pipeline/ gpt-4o-mini-2024-07-18
brand	Nordbrand Nordhausen	<u>Pilsner Urquell</u>	Pilsner Urquell
product category	Saure Kirsche	-	<u>Pilsner, Pils,</u> <u>Premuim</u> <u>Pilsner,</u> <u>Premium-Pils,</u> <u>Premium Pilsener,</u> <u>Pre-</u> <u>mium Pilsner,</u> <u>Premium</u> <u>Pils, Pilsener</u>
GTINs	[04400065420106]	-	[08594404000480, 08594404110110]
weight number	0.7	<u>0.33</u>	<u>0.33</u>
weight unit	Liter	<u>Liter</u>	<u>Liter</u>
different types	NaN	<u>NaN</u>	<u>NaN</u>
price	3.19	<u>3.99</u>	<u>3.99</u>
regular price	NaN	<u>3.99</u>	<u>0</u>
relative discount	NaN	<u>0</u>	<u>0</u>
absolute discount	NaN	<u>0</u>	<u>0</u>

Table 25: GT values and SOPs from the methods VLM/gemini-2.0-flash and Visual RAG Pipeline/gpt-4o-mini-2024-07-18 for the promotion image shown in Figure 6. Incorrect predictions are underlined.

Targets	GT	VLM/gemini-2.0-flash	Visual RAG Pipeline/ gpt-4o-mini-2024-07-18
brand	Fa	Fa	Fa
product category	Deo Roll-On, Roll-on, Deo-Roll on	-	<u>Deodorant,</u> Roll-on
GTINs	[04015000914246, 04015000914338]	-	[<u>1234567890123</u>]
weight number	50.0	50.0	50.0
weight unit	Milliliter	Milliliter	Milliliter
different types	yes	<u>no</u>	yes
price	0.88	0.88	0.88
regular price	1.35	1.35	1.35
relative discount	34	34	<u>35</u>
absolute discount	NaN	<u>0.47</u>	<u>0.47</u>

Table 26: VLM-based With/Without Fine-Tuning – Scores of the evaluation metrics for each single target variable and VLM. Percentages are calculated based on comparable GT and predicted values.

	gpt-4o- 2024-08-06	gpt-4o- 2024-08-06_FT	gemma-3-4b	gemma-3-4b- local_FT
brand	96.26%	<u>95.81%</u>	84.94%	93.40%
product category	-	6.84%	-	54.41%
GTINs	-	0%	-	5.60%
product weight	78.22%	<u>73.19%</u>	12.17%	61.36%
different types	90.81%	<u>87.50%</u>	69.20%	62.61%
price	97.83%	<u>92.16%</u>	69.19%	<u>96.70%</u>
regular price	<u>90.87%</u>	82.68%	38.10%	91.69%
relative discount	98.68%	<u>93.33%</u>	56.98%	0%
absolute discount	<u>88.76%</u>	<u>73.95%</u>	54.84%	95.49%

Table 27: VLM-based With/Without Fine-Tuning – Results of the evaluation metric \bigcup_{targets} for the structured output. The scores are calculated by systematically incorporating additional targets, based on the set of comparable GT values and predictions.

	gpt-4o- 2024-08-06	gpt-4o- 2024-08-06_FT	gemma-3-4b	gemma-3-4b- local_FT
brand	96.26%	<u>95.81%</u>	84.94%	93.40%
\bigcup_{targets} pr. w.	75.49%	<u>70.73%</u>	10.37%	57.37%
\bigcup_{targets} diff. t.	<u>74.86%</u>	69.75%	9.53%	78.14%
\bigcup_{targets} price	<u>73.05%</u>	64.06%	5.64%	75.48%
\bigcup_{targets} reg. p.	<u>67.99%</u>	55.97%	0%	68.78%
\bigcup_{targets} rel. d.	69.21%	<u>56.51%</u>	0%	0%
\bigcup_{targets} pr. c.	-	0%	-	0%
\bigcup_{targets} GTINs	-	0%	-	0%
\bigcup_{targets} abs. d.	0%	0%	0%	0%

Table 28: VLM-based With/Without Fine-Tuning – Results of the evaluation metric \bigcup_{test} for the structured output. Scores are calculated based on comparable GT and predictions values relative to the entire test dataset.

	gpt-4o- 2024-08-06	gpt-4o- 2024-08-06_FT	gemma-3-4b	gemma-3-4b- local_FT
brand	96.26%	<u>95.21%</u>	83.45%	93.40%
\bigcup_{test} pr. w.	73.00%	<u>67.96%</u>	9.86%	55.48%
\bigcup_{test} diff. t.	31.42%	<u>29.52%</u>	2.09%	18.68%
\bigcup_{test} price	30.64%	<u>27.09%</u>	1.24%	18.01%
\bigcup_{test} reg. p.	9.90%	<u>8.22%</u>	0%	5.45%
\bigcup_{test} rel. d.	5.14%	<u>4.26%</u>	0%	0%
\bigcup_{test} pr. c.	-	0%	-	0%
\bigcup_{test} GTINs	-	0%	-	0%
\bigcup_{test} abs. d.	0%	0%	0%	0%

Table 29: VLM-based With/Without Fine-Tuning – Illustration of the average elapsed time and costs per request as well as the total elapsed time and costs across all requests for each VLM.

	gpt-4o- 2024-08-06	gpt-4o- 2024-08-06_FT	gemma-3-4b	gemma-3-4b- local_FT
avg. elapsed time per req.[s]	<u>3.48</u>	4.68	1.86	17.01
total elapsed time (all req.)[h]	<u>35.3</u> (=1.5d)	47.5 (=1.98d)	18.7	172.8 (=7.2d)
avg. total costs per req.[USD]	0.0022	0.0024	0.0001 ²	<u>0.0009</u> ²
total costs (all req.)[USD]	80.99	88.68	3.51 ²	<u>32.42</u> ²



(a) Advertisement image with incorrect price prediction.

(b) Price prediction and historical data as well as the GT value.

	price	historical data of price
gpt-4o-mini-2024-07-18	1.49	1.49 / 1.49 / 1.49
gemini-2.0-flash	1.69	1.49 / 1.49 / 1.69
qwen2.5vl-32b	1.49	1.69 / 1.49 / 1.49
mistral-small3.1-24b	1.36	1.49 / 1.49 / 1.36
gemini-2.5-flash	1.51	1.69 / 1.49 / 1.49
GT	1.89	

Figure 18: Illustration of an advertisement image for which the Visual RAG based approach incorrectly predicts the target *price* (18a) and the price predictions and the price specification from historical data for each VLM alongside the GT value (18b).

(a) Advertisement image with multiple price specifications.

(b) Price prediction and the GT value.

	price
gpt-4o-mini-2024-07-18	1.99
gemini-2.0-flash	1.99
llava-34b	1.99
qwen2.5vl-32b	1.99
mistral-small3.1-24b	1.99
llama3.2-vision-11b	NaN
gemini-2.5-flash	1.99
gpt-5-mini-2025-08-07	1.99
GT	1.93

Figure 19: Illustration of an advertisement image with multiple price specifications (19a) and the price predictions for each VLM and the GT value (19b).



(a) Advertisement image with multiple references to the product weight.

(b) Product weight predictions and the GT value.

	product weight
gpt-4o-mini-2024-07-18	112 gram
gemini-2.0-flash	112 gram
llava-34b	100 gram
qwen2.5vl-32b	112 gram
mistral-small3.1-24b	112 gram
llama3.2-vision-11b	NaN
gemini-2.5-flash	112 gram
gpt-5-mini-2025-08-07	112 gram
GT	16 pieces

Figure 20: Illustration of an advertisement image that presents multiple references to the product weight within the product description text (20a) and the product weight predictions for each VLM and the GT value (20b).