

Neuro-Symbolic Interpretation of Laundry Care Labels via Schema-Based Reasoning for Inconsistency and Information Detection

Josiah Wiggins¹

Aman Adhikari¹

Sedat Ozer¹

Abstract—Laundry care labels convey essential garment handling instructions through standardized symbols and accompanying text. However, automatic interpretation remains challenging not only due to small icon sizes and ambiguous visual cues, but also because existing solutions primarily focus on standardized symbols and fail to reliably extract additional free-form textual instructions commonly present on care labels. In this paper, we propose a neuro-symbolic, multi-modal pipeline for robust laundry care label understanding that leverages vision-language models. In our approach, an object detector identifies laundry care symbols, while a vision-language model focuses on text and extracts both symbol-equivalent and free-form textual instructions. Both modalities are converted into a shared symbolic schema capturing actions, constraints, and parameters such as temperature and heat level. Using this representation, unlike previously proposed symbol-detection frameworks, our proposed framework explicitly reconciles symbol meanings with text, categorizing each symbol as consistent with, inconsistent with, or not mentioned in the accompanying text, and performs binary conflict detection. Our experimental results demonstrate the effectiveness of combining neural models with explicit symbolic logic for complete and reliable interpretation of real-world instructional content by means of vision-language models.

I. INTRODUCTION

Recent years have seen increasing interest in applying deep learning to fashion analysis [1] and to the automation of laundry-related tasks with minimal human intervention [2]. Such technologies have broad potential applications, including assistive systems for elderly users, semi- or fully-autonomous robotic appliances [3], smart home environments, and accessibility tools for visually impaired individuals. In these settings, the reliable interpretation of laundry care labels is essential, as incorrect handling instructions may damage garments, shorten their lifespan, or lead to unsafe behavior in automated laundering systems.

A care tag, also referred to as a care label, is a small attachment typically found inside garments that provides handling instructions through a combination of standardized symbols and textual descriptions. These symbols follow internationally defined conventions, while the accompanying text may appear in multiple languages depending on the garment's market. In addition to symbol-equivalent instructions (e.g., wash temperature or ironing level), care labels frequently include additional free-form textual guidance, such as “wash with similar colors,” “wash inside out,” or “do not use fabric softener.” Although these instructions are not represented

by standardized symbols, they convey important actionable information and are critical for correct garment care [4].

Most existing approaches to laundry care label understanding focus primarily on recognizing standardized symbols or mapping mostly pre-defined textual descriptions to predefined symbol categories. Early work such as [5] proposed text-based care recognition systems, while more recent studies have explored multi-modal pipelines that incorporate both textual and visual information [6]. These multi-modal frameworks typically adopt either a two-stage approach that detects the care tag region followed by symbol recognition or a two-branch architecture in which text and symbols are processed independently and merged at a later stage. While such approaches improve recognition accuracy for standardized care symbols, they are largely designed to predict symbol classes and do not explicitly address the extraction or preservation of additional free-form instructions. As a result, previously proposed laundry-care symbol detection based solutions often produce incomplete interpretations of care labels. Free-form text-based instructions that do not correspond to standardized symbols are frequently discarded, ignored, or incorrectly forced into symbolic categories. At the same time, using only text-based or only a vision-language model (VLM) based solution that attempts to infer care instructions holistically may overgeneralize, hallucinate [7] symbol meanings, or fail to distinguish between symbol-equivalent text and genuinely non-symbolic guidance. We define hallucination in this context as predicting care instructions that are not supported by either textual or symbolic evidence on the label, i.e., instructions that should not be included in the output. For practical deployment in many automated laundry appliances, assistive technologies, and smart home environments, it is, therefore, insufficient to recognize symbols alone. Instead, such systems should aim to recover all actionable information present on the tag and organize it in a structured and interpretable manner. As shown in Fig. 1, care labels may include symbol-only instructions, conflicting symbol-text pairs, and additional free-form textual guidance, highlighting the challenges of reliable automated interpretation. Because care labels may contain conflicting instructions, our framework detects symbol-text inconsistencies and alerts the user when the extracted care actions may be unreliable.

In this paper, we argue that complete laundry care label understanding requires explicit multi-modal reasoning over both visual symbols and textual instructions. We propose a neuro-symbolic and schema-based framework that combines classical object detection algorithms with a vision-language

¹Dept. of Electrical and Computer Engineering, California State Polytechnic University, Pomona, CA 91768, USA {jtwiggins, aadhikari, sozer}@cpp.edu

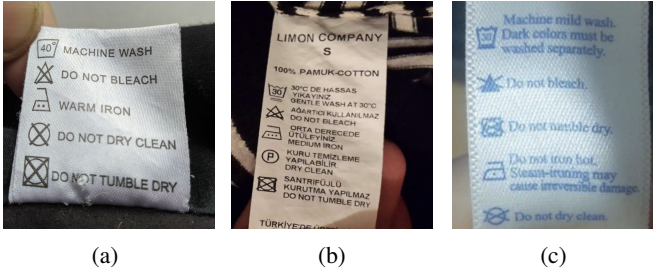


Fig. 1: Three sample images including care-labels. (a) laundry tag with symbol and symbol related text instructions; (b) laundry tag with inconsistent symbol and text (high iron symbol and medium iron text); (c) laundry tag including both symbol related text instructions and free-form instructions.

model (VLM) to extract and reconcile information from both modalities including image and text. By doing so, in addition to detecting the standard care symbols, we can also detect additional instructions and information, which does not have a standard format, provided on the tag. We evaluate our approach against multiple VLM-based baselines and demonstrate that deterministic schema-based reasoning substantially improves precision across all interpretation tasks, while enabling reliable extraction of free-form instructions, for which the previously proposed solutions were not designed to capture. Our main contributions are as follows:

- We introduce a multi-modal framework for care-label understanding on garments, where our framework jointly detects the (ISO) standardized laundry care symbols along with additional free-form textual instructions.
- We propose a symbolic schema based system that extracts structured fields and applies deterministic rules for a complete detection.
- We formulate multi-modal consistency, omission, and conflict detection as explicit reasoning tasks and empirically demonstrate that our proposed method performs better than multiple VLM-based baselines.

II. RELATED WORK

There has been significant interest in recent years in the automatic recognition of clothing and garment attributes from images for various fashion-related applications. Prior work in this area has focused on tasks such as material recognition, garment category classification, and edge or part detection [8], [9], [10], [11], [12]. These approaches primarily support fashion analysis and visualization applications, enabling users to preview how garments may appear on an individual or to organize large-scale fashion catalogs. While effective for high-level garment understanding, such methods do not address the interpretation of garment handling instructions conveyed through care labels. A related but comparatively underexplored line of research concerns the automatic recognition of laundry care instructions on garment tags, which is relevant to both fashion technology and smart home automation. Automatically interpreting care labels can enable robot-assisted or autonomous laundry systems to select appropriate

washing or drying settings without user intervention. Despite its practical importance, care label understanding remains a relatively novel application area, with only a limited number of studies reported in the literature [6]. Early work on this problem, such as [5], formulated care label interpretation primarily as a text and pictogram detection task. The authors introduced two small datasets consisting of 63 and 128 images, respectively, and employed a text-based E2E-MLT model [13] to jointly detect symbols and textual descriptions. Although symbol-only detection was briefly explored, the authors reported significantly lower performance compared to text-centric approaches. As a result, subsequent work has largely favored text-based or joint text-symbol detection pipelines. More recently, Su *et al.* [6] proposed a large-scale dataset, termed FashionCare, containing 157,907 care symbol images paired with corresponding textual instructions. Their proposed multi-modal multi-task framework (MMFC) adopts a two-branch architecture in which visual symbols and textual content are processed separately in the early stages before being combined. The framework employs EfficientNet for visual part and a Swin Transformer-based architecture for text processing. While this approach advances the scale and multi-modal nature of care label recognition, it remains focused on classifying standardized care symbols and their textual equivalents. Moreover, at the time of writing, the FashionCare dataset was not publicly available, limiting reproducibility and independent evaluation.

Symbolic reasoning has been widely studied as a means of enabling explicit and interpretable inference by representing actions and constraints in structured forms. Recent work has increasingly explored neuro-symbolic approaches that decouple neural perception from symbolic reasoning in order to improve logical consistency and faithfulness. Mao *et al.* proposed the Neuro-Symbolic Concept Learner, which maps visual detections to symbolic programs for compositional reasoning [14], while Yi *et al.* demonstrated improved reasoning fidelity by disentangling perception and reasoning in Neural-Symbolic VQA [15]. Constraint-aware symbolic reasoning has also been explored in physical and event-based domains, such as CLEVRER [16], and more recently through symbolic chain-of-thought formulations that improve the faithfulness of neural reasoning models [17]. However, symbolic reasoning has not been utilized for understanding laundry care tags. Importantly, prior care-label recognition approaches, including [5] and [6], primarily aim to detect standardized symbols or aligned text with predefined symbol categories. As a result, additional free-form textual instructions, such as “wash with similar colors,” “wash inside out,” or “do not use fabric softener”, are often ignored, discarded, or incorrectly mapped to symbolic classes. Furthermore, prior related work assumes consistency between visual symbols and provided textual descriptions, therefore, they do not explicitly address the cases in which the two modalities contradict each other or where symbol meanings are not included in the text entirely.

In contrast to previous work, our approach is designed to extract all actionable information present on a care label,

Symbol													
Classes	30C wash	40C wash	50C wash	60C wash	70C Wash	95C Wash	DN bleach	DN dry	DN dry clean	DN iron	DN steam	DN tumble dry	DN wash
Count	3950	3319	333	340	207	335	8213	95	4839	2421	112	5814	1221
Symbol													
Classes	bleach	chlorine bleach	drip dry	dry clean	dry clean any solvent	dry clean any solvent except trichloroethylene	dry clean low heat	dry clean no steam	dry clean petrol only	dry clean reduced moisture	dry clean short cycle	dry flat	hand wash
Count	187	89	360	156	174	3262	89	72	270	72	82	862	1139
Symbol													
Classes	line dry	line dry in shade	machine wash delicate	machine wash normal	natural dry	machine wash permanent press	non chlorine bleach	shade dry	tumble dry high	tumble dry low	tumble dry medium	tumble dry no heat	tumble dry normal
Count	443	96	177	236	48	137	641	238	48	1348	571	72	565
Symbol													
Classes	DN wet clean	DN wring	drip dry in shade	steam	iron	iron high	iron low	iron medium	wet clean	wring			
Count	49	294	8	39	104	409	4015	2467	242	46			

Fig. 2: This figure shows 49 laundry symbols and their meanings as classes as defined in the ISO-3758 standard. This table also lists the total number of instances per class in our dataset (see Section 4).

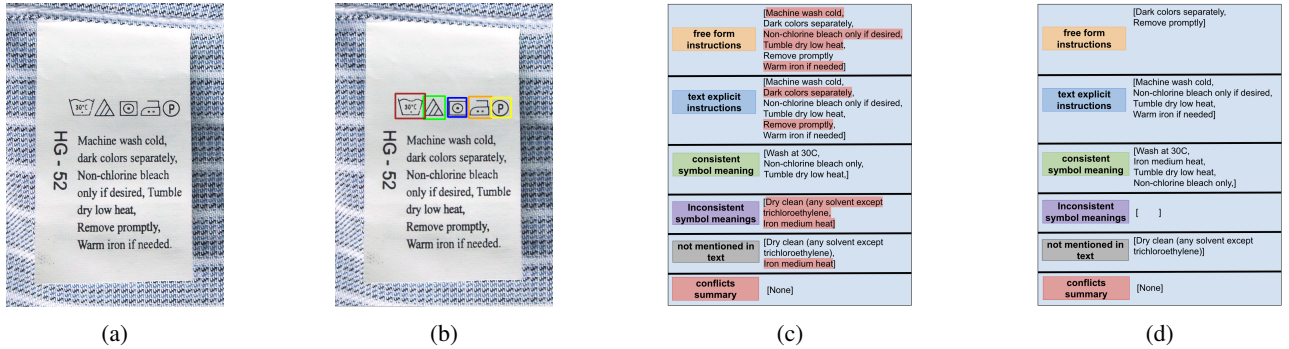


Fig. 3: End-to-end example. (a) Original care label image. (b) YOLOv11-large symbol detections. (c) VLM-only (Qwen 2.5) output. The outputs highlighted in red are hallucinated (i.e., incorrectly predicted). (d) Expected output, which is also the result obtained by our proposed end-to-end framework.

rather than focusing solely on standardized symbols. By converting both detected symbols and extracted text into a shared symbolic schema, our framework explicitly distinguishes symbol-equivalent instructions from non-symbol free-form guidance and enables deterministic reasoning over multi-modal content. This formulation allows consistency, omission, and conflict detection between symbols and text to be treated as core reasoning tasks, addressing limitations that are not explicitly considered in prior symbol-centric or text-centric care label recognition systems. Fig. 3 illustrates a representative end-to-end example, highlighting the limitations of text-centric vision-language inference and demonstrating how our proposed framework resolves hallucinated and inconsistent predictions through schema-based reasoning.

III. PROPOSED METHOD

We propose a neuro-symbolic, multi-modal framework for comprehensive laundry care label understanding. The

proposed framework combines neural models for visual and textual perception with deterministic symbolic reasoning to extract all actionable information present on a care label. Unlike prior approaches that focus primarily on standardized symbol recognition, our method jointly handles symbols, symbol-equivalent text, and additional free-form textual instructions, and explicitly reasons about consistency and contradictions between modalities.

A. Proposed Framework

Given a single image of a garment care label, the pipeline proceeds in five stages: (i) symbol detection, (ii) text extraction, (iii) symbolic schema construction, (iv) schema-based reconciliation, and (v) conflict detection. First, an object detection algorithm identifies standardized laundry care symbols directly from the image by computing their bounding box information. Next, a vision-language model extracts textual instructions using the detected symbols as

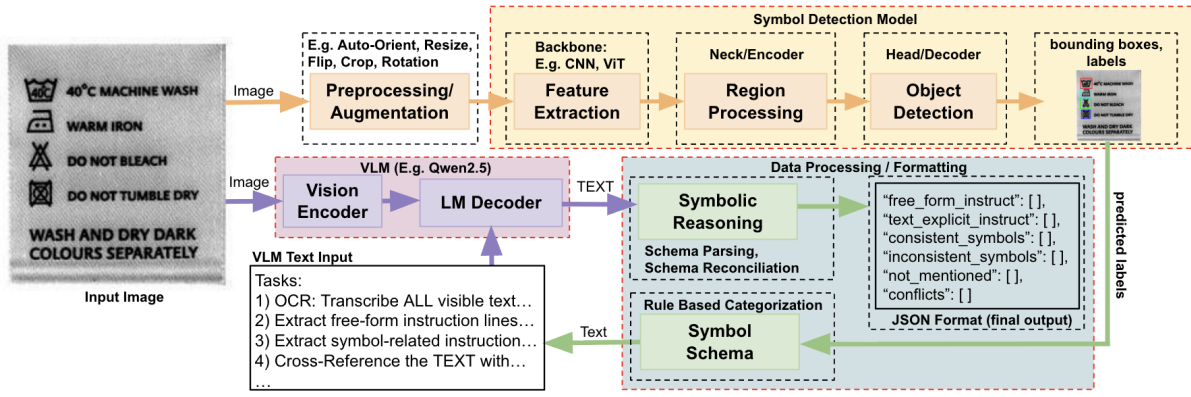


Fig. 4: Overview of the proposed laundry-care understanding framework. The system combines a symbol detection model, a vision-language model, and symbolic reasoning to produce normalized, machine-readable care instructions. The input to our framework is only a picture of the care label (no text-based input). The output of the framework includes bounding box information for each detected symbol and categorized instructions as text-based content.

context. Outputs from both modalities are mapped into a shared symbolic schema that encodes parameters and constraints about the additional instructions. These symbolic schema include: “action”, “prohibit”, “method”, “temp_c”, “level”, “max_level”, “solvent”, “modifier”, “source” and “evidence”. Deterministic reasoning over this schema categorizes symbol meanings as consistent with, inconsistent with, or not mentioned in the accompanying text, while preserving non-symbol free-form instructions. A binary conflict flag is produced when contradictions are detected. Figure 4 provides an overview of the pipeline.

Symbol Detection: Standardized laundry care symbols are detected using a state-of-the-art object detection model (such as YOLOv11 [18], YOLOv12 [19] or RF-DETR [20]) trained on symbol-level annotations. The detector operates directly on the input image without requiring a separate tag localization stage or multi-branch processing. This direct detection formulation reduces annotation requirements as it does not need text-based annotations and avoids cascading errors introduced by multi-stage pipelines. Each detected symbol is mapped to a standard semantic phrase (e.g., “wash at 40°C”, “do not tumble dry”). For stable downstream reasoning, duplicate detections of the same class are removed by class-level de-duplication. The resulting set of symbol meanings serves as the visual input to the symbolic reasoning stage.

Text Extraction and Instruction: Textual content on the care label is extracted using a vision-language model (VLM) capable of Optical Character Recognition (OCR) and instruction-level understanding. In our implementation, to leverage the power of VLMs, we employ the open-source, pre-trained Qwen2.5-VL-7B-Instruct model [21] for OCR and text extraction. However, the proposed reasoning framework is model-agnostic and can be applied to other vision-language models. Our approach uses existing pretrained VLM models, as they typically require large resources (including RAM and GPU) to re-train. Rather than fine-tuning the existing VLM model, we guide the existing model with a more structured prompt informed by the

outputs of the symbol detection stage. In our implementation, Qwen2.5-VL uses approximately 7 billion parameters, while a classic YOLO (YOLOv11-l in particular) algorithm has approximately 25 million parameters. Therefore, instead of re-training such a large VLM model, using the output of an additional object detection algorithm (such as YOLOv11) as input to the VLM is more computationally feasible, which is why we picked such an approach.

The text extraction module outputs *raw transcribed text* along with a set of *instruction lines*. These lines may include both symbol-equivalent (standard) instructions (e.g., *wash at 30°C*) and free-form guidance (e.g., *wash with similar colors*, *wash inside out*). Unlike prior implementations that attempt to map all text to predefined symbol categories, our approach explicitly preserves free-form instructions as additional and separate output, which does not directly associate with the detected symbols necessarily.

Symbolic Schema Representation: To enable explicit reasoning, both symbol-derived meanings (i.e., detected standard symbol-labels) and text-derived instructions (i.e., the output of the VLM algorithm) are converted into a shared symbolic schema. Each instruction is represented as a structured tuple encoding an action type (e.g., WASH, IRON, BLEACH), optional parameters such as temperature or heat level, and logical attributes including prohibition flags and constraints. Fig. 5 illustrates two representative examples of schema conversion, demonstrating how both symbol detections and textual instructions are mapped into a unified structured representation. For example, a textual instruction such as “do not iron hot” is represented as a constraint limiting the maximum allowable ironing level, whereas a symbol indicating iron high heat is represented as an explicit action with a specified level. This representation allows symbol meanings and textual instructions to be compared deterministically within a common semantic space. Free-form textual instructions that do not correspond to standardized symbols are retained separately and excluded from symbol reconciliation, ensuring that non-symbolic guidance is not

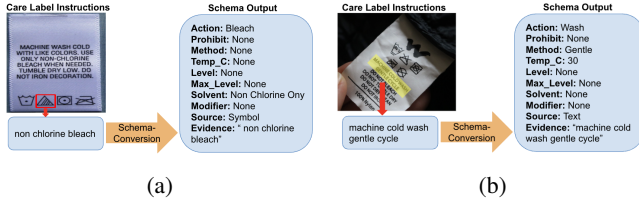


Fig. 5: Two examples of schema conversion from care label inputs to structured schema outputs. (a) shows an example for symbol-detection based conversion and (b) shows a text-extraction based conversion.

discarded or incorrectly forced into symbolic categories.

Schema-Based Reconciliation: Using the symbolic schema, the framework performs deterministic reconciliation between visual symbols and textual instructions. For each detected symbol meaning, the system evaluates its relationship to the extracted text and assigns it to one of three categories:

- Consistent: the symbol meaning is explicitly supported by the text;
- Inconsistent: the symbol meaning contradicts an explicit textual instruction or violates a textual constraint;
- Not mentioned: the symbol meaning is neither supported nor contradicted by the text.

Consistency is determined using strict matching rules that require agreement in action type and explicitly stated parameters. Inconsistencies are identified through logical contradictions, such as prohibit–allow conflicts or constraint violations (e.g., a symbol indicating high-heat ironing when the text specifies a maximum medium heat). Symbol meanings that lack textual counterparts are categorized as not mentioned.

Conflict Detection: In addition to symbol-level categorization, the system also performs binary conflict detection at the label level. A care label is marked as conflicting if at least one detected symbol meaning is classified as inconsistent with the accompanying text. This formulation allows conflict detection to be treated as an explicit and measurable task, enabling the system to flag contradictory instructions that may lead to unsafe or incorrect garment handling in automated or assistive applications.

IV. EXPERIMENTS

This section evaluates the proposed framework at two levels: (i) symbol detection performance, and (ii) end-to-end multi-modal interpretation and reasoning performance. The first set of experiments assesses the effectiveness of direct symbol detection using modern object detectors, while the second set evaluates the ability of the proposed neuro-symbolic pipeline to extract complete care label information, including free-form textual instructions and multi-modal inconsistencies. All the models were trained on an A100 GPU with 80GB VRAM, remotely.

Baselines and Comparisons: To evaluate the importance of explicit symbolic reasoning, we compare the proposed end-to-end pipeline against multiple baselines with different

TABLE I: A list of hyperparameters, as used in the experiments for all the YOLOv11, YOLOv12 and RF-DETR models, is provided.

Parameter	YOLO v11 & v12	RF-DETR
Epochs	250	250
Warmup epochs	3	0
Optimizer	AdamW	AdamW
Batch Size	64	16
Grad Accum Steps	–	1
Weight Decay	0.0005	0.0001
Image Size	640×640	varies*
Learning Rate	0.01	0.0001

Note: *RF-DETR input image sizes are 384×384 for the nano model, 512×512 for the small model, and 576×576 for the medium model.

levels of visual grounding and structured reasoning. Specifically, we consider a Qwen-only vision–language baseline, a Qwen + Detector baseline, and a LoRA Qwen + Detector baseline, in which Qwen2.5-VL-7B-Instruct is fine-tuned using Low-Rank Adaptation (LoRA) [22]. All methods are evaluated on the same dataset. For symbol detection, we evaluate multiple variants of YOLOv11, YOLOv12, and RF-DETR using standard object detection metrics. For the VLM component, we use Qwen2.5-VL-7B-Instruct [21]. For multi-modal reasoning, we compare all baselines and the proposed end-to-end pipeline using precision, recall, and F1 score.

All experiments are conducted on the curated laundry care label dataset described below. Images contain standardized laundry care symbols along with accompanying textual instructions. Ground truth includes symbol labels for free-form instructions, consistency, inconsistency, and omission between symbols and text. All detection models were trained for 250 epochs using AdamW. YOLOv11 and YOLOv12 used batch size of 64, image size of 640 × 640, weight decay of 0.0005, and learning rate of 0.01. RF-DETR used a batch size of 16, weight decay of 0.0001, learning rate of 0.0001, and model-specific input resolutions. Further training details for the used hyperparameters are listed in Table I.

Datasets: To support both symbol detection and end-to-end laundry care label understanding, we constructed a larger dataset by combining two publicly available care label datasets. The resulting larger dataset captures a wide range of real-world variability in care label appearance, including differences in lighting conditions, fabric textures, symbol occlusions, and background clutter. The first dataset, referred to as WL2¹, contains 1,794 images of laundry care labels. The second dataset, CareLabelsFind Computer Vision Model², consists of 3,251 images. Both datasets provide bounding-box annotations for standardized laundry care symbols but do not include tag-level localization or text annotations. Using the combined image pool, we applied a series of data augmentation strategies including rotations of up to ±15°, horizontal and vertical flipping, and random cropping (0%–20%). After augmentation, the final dataset comprises 8,043 images containing a total of 47,738 annotated symbol

¹<https://universe.roboflow.com/test-ah8ju/wl2>

²<https://universe.roboflow.com/carescan/carelabelsfind>

instances. The dataset is divided into training, validation, and test sets following a 70%/20%/10% ratio, which corresponds to 5,645 training images, 1,587 validation images, and 811 test images. The test set is used in all of our reported evaluations. Prior to training, all images undergo automatic orientation correction and are resized to a resolution of 640×640 pixels. In total, the dataset includes 49 distinct laundry care symbol classes based on the iso-3758 standard. Figure 2 provides a complete list of symbol classes along with the number of instances per class in the dataset. Only bounding-box annotations are available for each symbol. No textual or semantic annotations are provided, making the dataset suitable for evaluating direct symbol detection.

In addition to the symbol detection dataset, we constructed a separate dataset for training and evaluating the multi-modal reasoning baselines and the proposed end-to-end framework. This dataset consists of 616 care label images selected from the broader dataset and external sources to ensure the presence of both visual symbols and accompanying textual instructions. The images were chosen to cover a diverse range of real-world scenarios, including labels with consistent symbol-text pairs, omitted symbol mentions, contradictory instructions, and additional free-form guidance not represented by standardized symbols. To better simulate real-world conditions, we also applied data augmentation strategies including cropping, rotation, and noise injection. Unlike the symbol detection dataset, which contains only bounding-box annotations, this dataset was manually annotated at the label level to support multi-modal reasoning experiments. Ground truth annotations include (i) symbol presence, (ii) free-form textual instructions, (iii) symbol meanings consistent with text, (iv) inconsistent symbol meanings, and (v) symbol meanings not mentioned in text. These annotations enable quantitative evaluation of multi-modal interpretation, reconciliation, and conflict detection. The dataset is divided into training, validation, and test sets using a 67%/8%/25% split, corresponding to 414 training images, 52 validation images, and 150 test images. The training and validation splits are used for fine-tuning the LoRA Qwen + Detector baseline, while the held-out test set is used to evaluate all baselines as well as the proposed end-to-end framework under the same conditions. This setup enables a controlled comparison across methods while ensuring that evaluation is performed on unseen care label images. Out of these 616 images, approximately 18% of the reported care labels exhibit various forms of inconsistencies between symbolic and textual instructions, providing a meaningful subset for evaluating the framework’s ability to detect and reason about conflicting care instructions. Common forms of inconsistency observed in the dataset include mismatched ironing instructions, such as symbols indicating low-heat ironing while the accompanying text specifies medium heat, as well as dry-cleaning discrepancies where a symbol restricts solvents (e.g., any solvent except trichloroethylene) but the text provides only a generic “dry clean” instruction. While smaller in scale than the symbol detection dataset, this dataset is specifically designed to support both fine-tuning

TABLE II: Performance comparison across selected YOLOv11, YOLOv12, and RF-DETR variants on the symbol detection dataset.

Model	Prec	Recall	mAP ₅₀	Param (M)	Speed (ms)
YOLOv11-m	0.95	0.84	0.90	20.07	5.18 ± 0.20
YOLOv11-l	0.96	0.84	0.91	25.32	5.85 ± 0.25
YOLOv11-x	0.94	0.83	0.90	56.88	7.66 ± 0.15
YOLOv12-m	0.93	0.84	0.90	19.61	5.06 ± 0.20
YOLOv12-l	0.93	0.85	0.91	26.43	5.96 ± 0.05
YOLOv12-x	0.93	0.84	0.90	59.30	8.61 ± 0.07
RF-DETR-n	0.92	0.86	0.94	30.32	26.00 ± 0.00
RF-DETR-s	0.96	0.88	0.96	31.96	28.51 ± 0.00
RF-DETR-m	0.95	0.85	0.94	33.54	29.63 ± 0.00

and evaluation of multi-modal reasoning models and to expose failure cases that cannot be evaluated using symbol-only annotations.

Used Metrics: We evaluated models with multiple metrics including precision, recall, and mean Average Precision (mAP). Precision ($Prec$) is defined as $Prec = \frac{TP}{TP+FP}$, where TP is the number of true positives and FP is the number of false positives. Recall (R) is defined as $R = \frac{TP}{TP+FN}$, where FN is the number of false negatives. Average Precision (AP) is computed as the area under the precision-recall curve: $AP = \int_0^1 Prec(R) dR$. The mean Average Precision (mAP) is computed at a fixed IoU threshold of 0.50, denoted as mAP₅₀.

Symbol Detection Performance: Table II summarizes the symbol detection performance of multiple YOLOv11, YOLOv12, and RF-DETR variants. Performance is evaluated over multiple metrics including precision, recall, and mean Average Precision (mAP). Additionally, we report inference speed (mean and standard deviation in milliseconds) for each model. Across YOLOv11 and YOLOv12 variants, we observe improvements in precision and recall as model size increases, with YOLOv11-l achieving the strongest overall balance between mAP₅₀, precision and recall among the YOLO family. RF-DETR variants achieve the highest detection accuracy overall, with RF-DETR-s and RF-DETR-m reaching mAP₅₀ values of 0.96 and 0.94, respectively. However, this performance comes at a substantially higher computational cost, with inference speeds an order of magnitude slower than YOLO-based models.

Overall, these preliminary experimental results demonstrate that direct symbol detection without tag localization or needing text-based annotations is sufficient to achieve high and reliable detection accuracy, validating the design choice adopted in this work. Given the real-time constraints of smart home and robotic applications, YOLO-based models offer a favorable trade-off between accuracy and efficiency and, consequently, YOLOv11-l is used in our end-to-end experiments.

End-to-End Multi-modal Interpretation: While strong symbol detection is important, accurate care label understanding requires extracting and organizing all actionable

TABLE III: Comparison across instruction types in terms of averaged precision, recall, and F1 scores for Qwen-only, Qwen + Detector, LoRA Qwen + Detector, and the proposed End-to-End framework.

Instruction Type	Qwen-only			Qwen + Detector			LoRA Qwen + Detector			End-to-End		
	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Free-form	0.38	0.87	0.53	0.38	0.87	0.53	0.41	0.90	0.57	0.71	0.97	0.82
Consistent	0.35	0.12	0.18	0.47	0.87	0.61	0.50	0.87	0.65	0.96	0.87	0.91
Inconsistent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.88	0.70
Not mentioned	0.00	0.00	0.00	0.47	0.37	0.41	0.50	0.41	0.45	0.84	0.90	0.87
Symbol presence	0.46	0.40	0.43	0.87	0.91	0.89	0.87	0.91	0.89	0.87	0.91	0.89



Fig. 6: This figure shows sample results on three images. The first row shows the input images. The second row shows the detected symbols and their types (color-coded), and the third row shows parsed text-based instructions. For the second row, the following color scheme is used: red for 40C wash, green for DN bleach, blue for DN tumble dry, orange for iron low, teal for DN dry clean, brown for hand wash, and cyan for iron medium.

information on the tag. As Table III demonstrates, our proposed end-to-end framework substantially yielded higher performance in precision, recall and F1 scores. This demonstrates that explicit schema-based filtering and reconciliation are important for reliably extracting non-symbolic free-form instructions such as “wash with similar colors” or “wash inside out.” For symbol meanings consistent with text, performance improves substantially from Qwen-only to Qwen + Detector and LoRA Qwen + Detector, with the proposed end-

to-end framework achieving the best overall precision, recall, and F1 score. These gains highlight the benefit of explicit schema-based matching over implicit neural inference. The most pronounced difference appears in detecting inconsistent symbol meanings. Qwen-only, Qwen + Detector, and LoRA Qwen + Detector all fail entirely on this task, achieving zero precision, recall, and F1 score. For inconsistent symbol meanings, all VLM-based baselines fail, whereas the proposed framework achieves an F1 score of 0.70, highlighting

the importance of explicit symbolic reasoning for contradiction detection. For symbols not mentioned in text, the proposed method also outperforms all baselines with an F1 score of 0.87. For image-level symbol presence detection, all detector-based methods achieve the same F1 score of 0.89, indicating that the main advantage of the proposed framework lies in higher-level reasoning tasks rather than symbol grounding alone. Fig. 6 presents qualitative results of the proposed pipeline, including symbol detection, text extraction, and schema-based reconciliation.

V. CONCLUSION

In this paper, we presented a neuro-symbolic, multi-modal framework for comprehensive laundry care label understanding that goes beyond conventional symbol recognition and detection. By combining neural perception with explicit symbolic reasoning, our proposed framework is able to detect standardized care symbols, preserve additional free-form textual instructions, and explicitly reason about consistency and contradictions between visual and textual modalities. This formulation addresses a key limitation of prior work, which typically focuses on symbol classification or text-based inference while overlooking non-symbolic guidance and multi-modal inconsistencies commonly present on real-world care labels, as illustrated in Fig. 1.

Through our experiments, we first demonstrated that modern object detection models can accurately recognize laundry care symbols using a direct detection approach, without requiring multi-stage or multi-branch architectures. Then, we showed that explicit schema-based reasoning substantially improves performance on higher-level interpretation tasks. In particular, our proposed framework achieved the strongest results for free-form extraction, consistent instruction matching, omitted symbol reasoning, and inconsistent symbol detection. Most notably, all the tested VLM-based baselines failed to detect inconsistent symbol meanings, whereas the proposed framework successfully identified such contradictions with strong precision and recall. At the same time, while detector-based baselines matched the proposed system on symbol presence detection, they did not provide the reasoning capability needed for consistency, omission, and conflict analysis.

These findings highlight the importance of combining neural models with deterministic symbolic logic for reliable interpretation of instructional content in real-world settings. Beyond laundry care labels, the proposed framework is applicable to a broader class of multi-modal documents that contain structured symbols alongside unstructured text, such as safety labels, product manuals, and regulatory signage. Future work will explore extending the symbolic schema to additional domains, incorporating multilingual reasoning, and integrating the system into real-time assistive and robotic applications.

REFERENCES

- [1] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhaes, D. Goncalves, C. Greco, and J. Tagliabue, "Contrastive language and vision learning of general fashion concepts," *Scientific Reports*, vol. 12, no. 1, p. 18958, 2022.
- [2] P. Barnard, M. J. G. Trigo, D. Price, S. Cobb, G. Reyes-Cruz, G. Berumen, D. Branson III, M. A. Khanesar, M. T. Torres, and M. Valstar, "Human strategies for correcting human-robot errors during a laundry sorting task," *arXiv preprint arXiv:2504.08395*, 2025.
- [3] W. B. Bedada, I. Ahmadli, and G. Palli, "Robotized laundry manipulation with appliance user interface interpretation," in *Human-Friendly Robotics 2021: HFR: 14th International Workshop on Human-Friendly Robotics*. Springer, 2022, pp. 91–106.
- [4] A. K. Ramratan and M. Kumar, "Study on garments care symbols and laundry performance."
- [5] J. Kralicek, J. Matas, and M. Busta, "Care label recognition," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 959–966.
- [6] M. Su and Z. Wang, "Large scale multimodal fashion care recognition," *Multimedia Systems*, vol. 30, no. 5, p. 304, 2024.
- [7] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," *arXiv preprint arXiv:2509.04664*, 2025.
- [8] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.
- [9] S. Guo, W. Huang, X. Zhang, P. Srikantha, Y. Cui, Y. Li, H. Adam, M. R. Scott, and S. Belongie, "The imaterialist fashion attribute dataset," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.
- [10] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, J. Tagliabue *et al.*, "Fashionclip: Connecting language and images for product representations," 2022.
- [11] H. Tian, Y. Cao, and P. Mok, "Detr-based layered clothing segmentation and fine-grained attribute recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3535–3539.
- [12] A. Baldriati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 23 393–23 402.
- [13] M. Bušta, Y. Patel, and J. Matas, "E2e-mlt-an unconstrained end-to-end method for multi-language scene text," in *Computer Vision—ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*. Springer, 2019, pp. 127–143.
- [14] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," *arXiv preprint arXiv:1904.12584*, 2019.
- [15] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding," *Advances in neural information processing systems*, vol. 31, 2018.
- [16] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clever: Collision events for video representation and reasoning," *arXiv preprint arXiv:1910.01442*, 2019.
- [17] J. Xu, H. Fei, L. Pan, Q. Liu, M.-L. Lee, and W. Hsu, "Faithful logical reasoning via symbolic chain-of-thought," *arXiv preprint arXiv:2405.18357*, 2024.
- [18] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [19] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.
- [20] I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri, "Rf-detr: Neural architecture search for real-time detection transformers," 2025. [Online]. Available: <https://arxiv.org/abs/2511.09554>
- [21] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.
- [22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," *Iclr*, vol. 1, no. 2, p. 3, 2022.

[1] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhaes, D. Goncalves, C. Greco, and J. Tagliabue, "Contrastive language and