Prompt-Calibrated Foundation-Model Fusion for RF20-VL Few-Shot Object Detection

Pu Luo[1] (25171214094@stu.xidian.edu.cn)
Cong Xu[1] (25171214052@stu.xidian.edu.cn)
Yumei Li[1] (25171213938@stu.xidian.edu.cn)
Licheng Jiao[1] (lchjiao@mail.xidian.edu.cn)
Puhua Chen[1] (phchen@xidian.edu.cn)
Dan Zhang[1] (see2011dan@126.com)
[1]National 111 project Base of Intelligent Information Processing, Xi'an, China

Abstract

This report describes our technical solution for the CVPR 2026 Roboflow-20VL Few-Shot Object Detection Challenge, In-Context Prompting Track. The method uses only pre-trained foundation models together with inference-time prompting, image preprocessing, candidate refinement, and post-processing. No gradient-based fine-tuning or parameter update is performed. We formulate each dataset as a small visual annotation problem: few-shot examples and category names are converted into dataset-specific class cues, multiple foundation-model detectors are prompted to generate category-conditioned candidates, and all candidates are normalized to a unified detection format. The final predictions are obtained by source-specific score calibration, coordinate-system recovery, geometry correction, class-aware or class-agnostic non-maximum suppression, and dataset-specific category remapping. The system combines Qwen-VL family models, Gemini vision models, Doubao vision models, SAM3 proposal/refinement, and classical post-processing, which together provide complementary semantic localization, independent visual priors, visual proposal quality, and robust submission formatting across diverse domains including document layout, industrial defects, medical imagery, products, aerial scenes, smoke, sports, and water-meter reading. The source code for this solution is available at https://github.com/lapuuu-810/Roboflow-20VL-Few-Shot-lababa.git.

1 Introduction

The RF20-VL benchmark evaluates few-shot object detection under a highly heterogeneous collection of visual domains. Each dataset contains only a small number of annotated examples but may differ substantially in object scale, appearance, imaging modality, category granularity, and annotation convention. The challenge is therefore not only to recognize the target categories, but also to adapt the localization procedure to each dataset without training a task-specific detector.

A direct prompt to a single vision-language model is often unstable in this setting. General visual localization models may miss small objects, confuse visually similar categories, produce loose boxes, or return coordinates in an inconsistent convention. Moreover, several datasets require domain-specific handling, such as contrast enhancement for X-ray images, panel extraction for water meters, mask-based refinement for irregular objects, or category-index remapping for Roboflow exports containing a dummy `none` class.

We address these issues with Prompt-Calibrated Foundation-Model Fusion. The central idea is to keep all adaptation at inference time. Few-shot examples and category names are used to construct compact visual and textual class descriptions. Several complementary foundation-model signals are then obtained and fused after explicit coordinate normalization and confidence calibration. Figure 1 summarizes the full technical route.

2 Challenge Protocol and Compliance

Our solution follows the In-Context Prompting Track constraint. All model adaptation is performed through prompts, visual examples, preprocessing, candidate selection, and post-processing. We do not update model weights, train a neural detector, tune gradients, or use test-set labels during inference. The few-shot annotations are used only as in-context visual evidence and for local validation of prompt variants, threshold choices, score multipliers, box scaling, and fusion policies.

For each dataset, the final output is a pickle file containing one record per image. Each record stores the image identifier and a list of detected instances, where each instance contains `image_id`, `category_id`, `bbox`, and `score`. All submitted boxes are represented in the required $[x, y, w, h]$ format in the original image coordinate system.

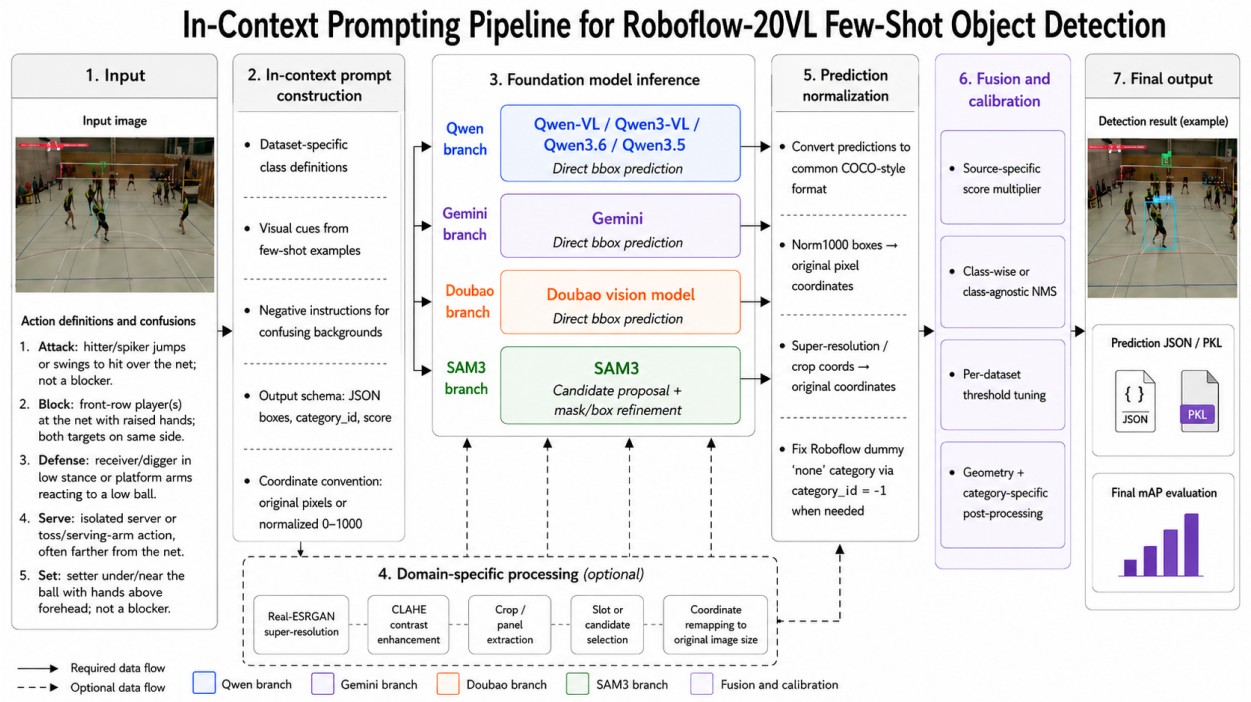

Figure 1: Overview of the proposed in-context prompting detection pipeline. The method first builds task-specific prompts with class definitions, visual cues, negative constraints, output schemas, and coordinate rules. Multiple foundation-model branches, including Qwen, Gemini, Doubao, and SAM3, then perform direct bbox prediction or candidate refinement, optionally assisted by domain-specific preprocessing. All predictions are normalized to a unified COCO-style format and further refined by score calibration, NMS, threshold tuning, and geometry/category-aware post-processing.

3 Method

3.1 Overall Pipeline

Our framework adopts an in-context prompting based multi-branch detection pipeline. The example images in the figure are only used to illustrate the detection scenario, while the core method focuses on prompt construction, foundation-model inference, domain-specific processing, prediction normalization, and calibrated fusion.

First, task-specific prompts are constructed by integrating class definitions, few-shot visual cues, negative confusion rules, structured output requirements, and coordinate conventions. The class definitions emphasize discriminative visual patterns of each target category, while the negative instructions are used to reduce common misclassifications among visually similar actions or background regions. The output schema further constrains the vision-language models to produce detection-style predictions, including bounding boxes, category identifiers, and confidence scores.

Based on the constructed prompts, multiple complementary inference branches are employed. Qwen-series models, Gemini, and the Doubao vision model are used for direct bounding-box prediction, providing prompt-guided localization results from different foundation-model priors. In parallel, SAM3 serves as a proposal-oriented branch for candidate generation and mask/box refinement, offering an additional localization prior that complements direct predictions from vision-language models.

To improve robustness under challenging visual conditions, an optional domain-specific processing branch is introduced. This branch includes operations such as super-resolution, contrast enhancement, crop or panel extraction, candidate selection, and coordinate remapping. These modules enhance local visual details, reduce background interference, and ensure that predictions generated on processed regions remain spatially consistent with the original image space.

After inference, predictions from all branches are normalized into a unified COCO-style format. This step converts normalized coordinates to original pixel coordinates, remaps boxes from cropped or super-resolved regions, and handles dataset-specific label offsets when necessary. Finally, the normalized predictions are fused through source-specific score calibration, class-wise or class-agnostic NMS, per-dataset threshold tuning, and geometry/category-aware post-processing. This fusion stage suppresses redundant detections, calibrates confidence scores across different sources, and produces the final predictions for evaluation.

3.2 Dataset-Specific Prompt Construction

Generic prompts are insufficient for RF20-VL because the datasets include medical X-ray images, industrial textures, document pages, shelf products, aerial scenes, and structured

reading panels. We therefore manually adapt prompts from the few-shot examples and class labels. Each prompt contains four elements:

1. concise category definitions and visually discriminative cues;
2. negative constraints that suppress background, texture, panel borders, shadows, and confusing non-target objects;
3. a coordinate convention, either normalized 0-1000 coordinates or original-pixel boxes;
4. a JSON-only output schema containing category name or id, bounding box, and confidence.

For datasets with repeated structured regions or very small objects, we use a two-stage prompting strategy. The first stage localizes coarse regions or candidate sets, while the second stage classifies or selects candidates within those regions.

3.3 Foundation-Model Evidence

Qwen-VL family models. Qwen-VL variants are used as the primary semantic localization models. They are prompted with dataset-specific class descriptions and few-shot visual cues, and return category-conditioned boxes in a strict schema. For document-layout and UI-like datasets, normalized 0-1000 coordinate prediction is especially convenient and is later rescaled to the original image size.

Gemini vision models. Gemini is introduced as an additional direct bounding-box prediction branch. Compared with relying on a single VLM family, Gemini provides an independent semantic and visual prior, which is useful for recovering missed targets and improving robustness under domain shifts. In practice, Gemini predictions are treated as an independent source and are normalized, calibrated, and fused with other model outputs in the same way as Qwen and Doubao predictions.

Doubao vision models. Doubao is used as another independent VLM detector. It often contributes complementary recall when Qwen or Gemini misses salient objects, or when a domain benefits from a different visual prior. We usually assign Doubao a dataset-specific source multiplier before fusion.

SAM3. SAM3 is used as a proposal generator, mask-to-box refinement module, panel/slot selector, and object-shape filter. In several datasets, VLMs provide semantic labels or coarse locations, while SAM3 supplies boundary-aware boxes or masks. This is particularly useful for products, defects, bottles, dreidels, water meters, and low-contrast X-ray regions.

3.4 Coordinate Recovery and Box Calibration

All predictions are normalized to a unified representation before fusion. If a model outputs normalized coordinates $(x_1, y_1, x_2, y_2) \in [0, 1000]^4$, we map them to the original image size (W, H) as

$$b = \left[\frac{x_1 W}{1000}, \frac{y_1 H}{1000}, \frac{x_2 W}{1000}, \frac{y_2 H}{1000}\right]. \tag{1}$$

For crop, tile, slot, or super-resolution pipelines, the inverse transform is applied so that every candidate box is expressed in the original image coordinate system before evaluation or submission.

Some VLMs produce boxes that are systematically tight or loose. We therefore apply a center-fixed scaling operation when local validation shows a consistent bias. Let a box have center (c_x, c_y), width w, height h, and scale factor γ. The calibrated box is

$$b_\gamma = \left[c_x - \frac{\gamma w}{2}, c_y - \frac{\gamma h}{2}, c_x + \frac{\gamma w}{2}, c_y + \frac{\gamma h}{2}\right]. \tag{2}$$

The scale factor is selected from a small local sweep and clipped to image boundaries.

3.5 Late Fusion and Filtering

Each prediction source is assigned a dataset-specific score multiplier. For a candidate with original score s_i from source s, the calibrated confidence is

$$s_i' = \alpha_s \cdot s_i, \tag{3}$$

where α_s is selected according to local validation behavior. Candidates are then merged with either class-aware or class-agnostic non-maximum suppression. Class-aware NMS is the default for multi-class datasets, while class-agnostic NMS is used when duplicated boxes across categories are more harmful than class confusion.

Additional filters are applied only when they reflect stable dataset priors, such as plausible area ranges, aspect-ratio constraints, product-grid regularity, smoke-region geometry, or protection rules for confusing age-like categories. These filters are deterministic and do not introduce learned parameters.

4 Dataset-Specific Strategies

Table 1 summarizes the main specialization used for each dataset. The purpose of these strategies is not to train separate detectors, but to adapt the inference pipeline to the visual structure and annotation convention of each dataset.

5 Implementation Details

5.1 Prediction Schema

All detector outputs are first converted to a common internal schema containing the image id, category id, source name, $[x_1, y_1, x_2, y_2]$ box, confidence score, and optional auxiliary metadata such as crop id, tile offset, mask area, or prompt

Table 1: Dataset-specific inference strategies used in the final system.

Dataset	Main evidence sources	Key adaptation and post-processing
actions-zzid2-zb1hq-fsod-amih	Qwen action-oriented VLM pipeline	Use action/person/object relationship prompts and convert VLM outputs into detection instances.
aerial-airport-7ap9o-fsod-ddgc	VLM detections + geometric priors	Apply box calibration, area regularization, and filtering of implausible aerial detections.
all-elements-fsod-mebv	Qwen + Doubao	Detect UI elements with class-specific prompts; fuse Doubao as a low-weight recall source and apply class-wise NMS.
aquarium-combined-fsod-gjvb	Qwen + Doubao	Fuse Qwen baseline with Doubao recall detections; remap local category ids by removing the Roboflow dummy class.
defect-detection-yjplx-fxobh-fsod-amdi	Qwen + Doubao	Use multi-pass VLM defect detection; retain useful Doubao defect classes and fuse with Qwen predictions under high NMS tolerance.
dentalai-i4clz-fsod-fsuo	Qwen + Gemini + SAM3 crop candidates	Generate SAM3 crop candidates and classify dental findings with Qwen; remap local ids to server ids after removing the dummy class.
flir-camera-objects-fsod-tdqp	SAM3 initial detections	Convert SAM3 boxes to COCO format; tune score threshold, NMS, top-k filtering, and center-fixed box scaling.
gwhd2021-fsod-atsv	SAM3 initial detections	Use dense SAM3 wheat-head proposals; apply class-wise NMS, per-image top-k filtering, and box scale calibration.
lacrosse-object-detection-fsod-uxkt	Multiple VLM variants + fallback	Fuse several VLM outputs and apply mild center-fixed box scaling for small sports objects.
new-defects-in-wood-uewd1-fsod-tffp	Multiple VLM variants + fallback	Use fine-grained prompts for cracks, holes, knots, and knot cracks; fuse multiple VLM outputs with fallback candidates.
orionproducts-vtl2z-fsod-puhv	Qwen + Gemini + SAM3	Localize repeated product/SKU instances with SAM3 proposal support; use low thresholds and NMS for dense shelf scenes.
paper-parts-fsod-rmrg	Qwen direct layout detector	Predict normalized document-layout boxes and rescale them back to original image coordinates.
recode-waste-czvmg-fsod-yxsw	Doubao + Qwen tile-filtered	Use Doubao as the base detector and add Qwen tiled detections for small or partially occluded waste objects.
soda-bottles-fsod-haga	SAM3 color prompts + Doubao	Use color/shape prompts to distinguish bottle categories; fuse SAM3 detections with Doubao complements.
the-dreidel-project-anzyr-fsod-zejm	SAM candidates + Qwen crop classification	Generate object candidates with SAM and classify crops with Qwen; stabilize type/symbol assignments with bias post-processing.
trail-camera-fsod-egos	SAM3 initial detections	Convert SAM3 animal proposals to final detections; tune NMS, top-k filtering, and box enlargement for deer/hog instances.
water-meter-jbktv-7vz5k-fsod-ftoz	Qwen panel localization + SAM3 slots	Treat the task as structured panel reading; convert digit/slot candidates into detection boxes and map them back to the image.
wb-prova-stqnm-fsod-rbvg	Qwen + SAM refinement + prompt/KNN cue	Refine boundaries with SAM, apply center-fixed scaling, and use category-protection rules to reduce class confusion.
wildfire-smoke-fsod-myxt	Doubao + Qwen	Use Doubao for broad smoke regions and Qwen for recall; apply offset and scale calibration for amorphous smoke boxes.
x-ray-id-zfisb-fsod-dyjv	Enhancement + SAM3 + geometry	Use Real-ESRGAN/CLAHE for contrast enhancement; scale predictions from enhanced images back to the original resolution.

variant. This unified schema makes it possible to share fusion and validation code across datasets even when the raw model outputs differ.

The final submission for each dataset follows the required pickle structure: a list of image records, each with an `image_id` and an `instances` list. Each instance contains `image_id`, `category_id`, a `bbox` in $[x, y, w, h]$ format, and a floating-point `score`.

For datasets exported from Roboflow with a dummy `none` class at `category_id=0`, we subtract one from each real-class id before submission, because the server convention removes `none` and uses zero-based ids for valid categories.

5.2 Local Model Selection and Parameter Sweeps

The method uses small local sweeps over score thresholds, source multipliers, NMS thresholds, and box scale factors. These sweeps are performed on the available local annotations and are used only to select deterministic inference hyperparameters. Common sweep variables include α_s for each prediction source, the NMS IoU threshold, the minimum confidence threshold, and the center-fixed scale factor γ. The final selected configuration is then applied to all test images in that dataset.

5.3 Robustness Considerations

Several implementation details are important for stability. First, all coordinate transformations are logged and checked because normalized boxes, crop boxes, tile boxes, and super-resolution boxes can otherwise be mixed accidentally. Second, prediction fusion is performed after category-id normalization, not before. Third, low-coverage results are not used as final submissions even when they achieve strong local performance on a subset, because full image coverage is required. Fourth, dataset-specific rules are kept conservative and deterministic, so that they correct systematic errors without becoming a hidden training procedure.

6 Discussion

The proposed system benefits from the complementarity among different foundation models and processing strategies. Qwen-VL family models provide strong semantic grounding and are effective when class names, textual definitions, and few-shot visual cues clearly describe the target object or action. Gemini is introduced as an additional direct prediction branch, offering an independent visual-semantic prior that improves robustness when predictions from one VLM family are incomplete or unstable. Doubao further improves recall in several datasets by providing another independent visual prior, which helps recover detections missed by the Qwen or Gemini branches. SAM3 improves proposal quality and boundary alignment, especially when objects have clear contours or when localization can be refined from masks, slots, or candidate regions. Therefore, the final performance does not rely on a single model, but on the combination of multiple semantic prediction branches, independent visual priors, and proposal-level refinement.

In addition to model complementarity, data-specific preprocessing plays an important role in improving detection robustness. Since the involved datasets vary significantly in image resolution, object scale, contrast, texture, and background complexity, a fixed preprocessing strategy is often suboptimal. For small or low-resolution targets, super-resolution methods such as Real-ESRGAN, as well as simple linear interpolation, can enlarge local regions and make fine-grained visual cues more recognizable to vision-language models. This is particularly useful when the model needs to detect small objects or actions after a coarse-to-fine localization process. In our pipeline, a large region can first be localized, and then a smaller crop is used for refined detection. However, when the cropped image becomes too small, models that return bounding boxes in a normalized 0–1000 coordinate system may produce unstable or shifted boxes. Directly using the original full image can also be problematic, because the model may attend to irrelevant regions and generate false positives. Enlarging the crop before inference alleviates this issue by preserving the target location while providing sufficient spatial resolution for more accurate box prediction.

Contrast and appearance enhancement are also beneficial for specific visual domains. For example, CLAHE can improve local contrast and highlight structural details in low-contrast images, which is especially helpful for X-ray-like data or images where object boundaries are weak. Similarly, crop or panel extraction can reduce background interference when the target region only occupies a small part of the image. These observations suggest that in-context prompting based detection is not only sensitive to prompt design, but also to the visual quality and spatial scale of the input provided to the model. Carefully designed domain-specific preprocessing can therefore provide a meaningful performance gain without introducing any gradient-based training.

Another important factor is reliable coordinate handling. Different branches may produce predictions under different coordinate systems, including original image coordinates, normalized 0–1000 coordinates, super-resolved coordinates, and cropped-region coordinates. If these predictions are not accurately mapped back to the original image space, even semantically correct detections may lead to poor localization scores. Therefore, prediction normalization and coordinate remapping are essential components of the pipeline. This is especially important for crop-based refinement, where small coordinate errors in the crop space can become large offsets after remapping.

The main limitation of the pipeline is that it remains dataset-specific. Prompt design, preprocessing choices, score calibration, threshold selection, and geometry correction must be adjusted for different visual domains. However, this is a natural trade-off under the In-Context Prompting Track. Without gradient-based training or model fine-tuning, performance mainly comes from careful prompt engineering, domain-aware visual enhancement, reliable coordinate conversion, and robust multi-source fusion rather than a single learned detector. In this setting, dataset-specific adaptation is not merely a limitation, but also an effective way to exploit the flexibility of foundation models under few-shot constraints.

7 Conclusion

We presented Prompt-Calibrated Foundation-Model Fusion for the CVPR 2026 Roboflow-20VL Few-Shot Object Detection Challenge. The method treats each dataset as a small in-context visual annotation task and combines Qwen-VL family models, Gemini vision models, Doubao vision models, SAM3 candidate refinement, preprocessing, coordinate recovery, confidence calibration, late fusion, and deterministic submission remapping. The complete pipeline remains within the inference-time adaptation setting because it uses no gradient-based fine-tuning. This design provides a practical and robust solution for heterogeneous few-shot object detection datasets with diverse imaging domains and annotation conventions.