GENERALIZING VLMS TO EXTREME DOMAINS: BENCHMARKING, REWARD-DRIVEN POST-TRAINING, AND SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern vision-language models (VLMs) appear general because a single autoregressive interface can solve many visual tasks. Yet task breadth does not imply robustness across visual worlds: current evidence is drawn largely from natural imagery resembling web-scale pretraining data. We investigate whether unified VLM perception survives extreme shifts in image formation and visual statistics. To make this question measurable, we introduce CrossVLM-Bench, a benchmark for both post-training and evaluation across remote sensing, industrial inspection, agriculture, and underwater imagery. It contains 7,362 images and 66,258 image–task pairs, with every image supporting nine tasks spanning recognition, localization, segmentation, grounding, reasoning, and region-level description. Its strictly image-disjoint splits combine human-annotated source geometry with human-verified task and language annotations. Evaluation across four VLM families reveals a systematic breakdown that extends beyond recognition to spatial precision, object-set completeness, and language–region grounding. Target-domain supervised fine-tuning (SFT) partially recovers these capabilities, but optimizes the likelihood of a particular response serialization rather than the correctness of the decoded perceptual outcome. We therefore propose CD-GRPO, a reward-driven post-training framework that recasts extreme-domain adaptation as structured outcome optimization. CD-GRPO verifies complete decoded predictions using task-native semantic and geometric criteria while enforcing valid, complete, and nonredundant output structure. On Qwen3-VL-8B, CD-GRPO raises the average score across four domains and nine tasks from 25.7 to 48.5, exceeding SFT by 6.8 points and the strongest recent perception-oriented RL baseline by 7.7 points under equivalent protocols. Scaling analyses across model capacity, target-domain data, and test-time adaptation further characterize when additional resources translate into domain robustness. Together, our results establish extreme-domain robustness as a distinct axis of VLM generality and structured outcome optimization as an effective route toward it. We will release the benchmark, model, and code.

1 INTRODUCTION

Modern vision-language models (VLMs) are evolving toward general-purpose visual systems capable of recognizing, localizing, segmenting, counting, reasoning about, and describing visual content through a shared interface Bai et al. (2025); Wang et al. (2025); Xiaomi et al. (2025). However, task breadth and domain robustness are distinct dimensions of generality. Current evidence is derived predominantly from natural-image domains, whereas many consequential applications operate in visual regimes that differ fundamentally from web-scale imagery. Remote sensing Li et al. (2020), industrial inspection Song & Yan (2013), precision agriculture Wei et al. (2024), and underwater exploration Jiang et al. (2021) exhibit substantial shifts in image formation, viewpoint, scale, texture, illumination, and object density. Whether the unified perceptual competence of modern VLMs survives such extreme domain shifts remains largely unanswered.

Existing evaluation paradigms do not adequately expose this problem. General-purpose VLM benchmarks Yu et al. (2016); Lai et al. (2024) expand the range of tasks while largely retaining

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

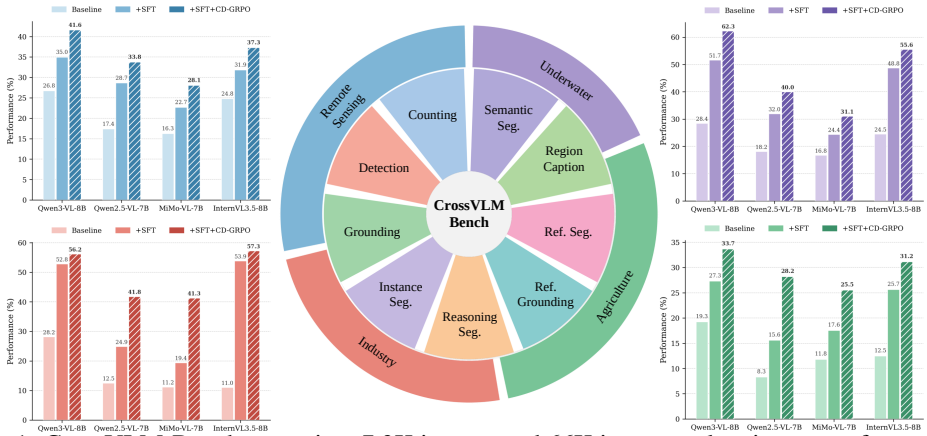

Figure 1: CrossVLM-Bench comprises 7.3K images and 66K image–task pairs across four extreme visual domains: remote sensing, industrial inspection, agriculture, and underwater imagery. Each image supports all nine perception tasks, with strictly image-disjoint training and test splits used for post-training and evaluation, respectively. We report the mean performance across the nine tasks for four representative VLMs before adaptation, after SFT, and after SFT followed by CD-GRPO. All models exhibit substantial performance degradation in these domains, while CD-GRPO consistently achieves the strongest adaptation performance across models and domains.

natural-image distributions, whereas cross-domain benchmarks Triantafillou et al.; Guo et al. (2020); Lei et al. (2022); Fu et al. (2025) vary the visual distribution but typically evaluate isolated tasks using specialized vision models. This leaves a critical intersection unexplored: how domain shift simultaneously affects the diverse capabilities unified within a VLM. Evaluating detection, segmentation, grounding, counting, and reasoning independently cannot reveal whether these failures share common causes or whether they can be recovered through a unified adaptation framework.

To address this gap, we introduce CrossVLM-Bench, a unified multi-task benchmark with dedicated training and test splits for VLM post-training and evaluation. It comprises 7,362 images from four extreme visual domains: remote sensing, industrial inspection, agriculture, and underwater imagery. Every image supports nine perception tasks: detection, semantic segmentation, instance segmentation, counting, object grounding, referring grounding, referring segmentation, reasoning segmentation, and region captioning. The resulting 66,258 image–task pairs integrate semantic, geometric, and linguistic targets on the same visual evidence, supporting multi-task post-training and holistic evaluation of domain robustness within a common task space. CrossVLM-Bench is grounded in reliable annotation provenance. Categories, bounding boxes, and available masks are inherited from human-annotated source datasets, and geometry-dependent targets are derived deterministically. GPT-4o Islam & Moushi (2025) assists in converting the source annotations into standardized instruction–response pairs for the nine tasks and generating the associated language content; all GPT-assisted annotations are subsequently verified and corrected by trained human annotators.

Building on CrossVLM-Bench, we evaluate Qwen2.5-VL Yang et al. (2025), Qwen3-VL Bai et al. (2025), InternVL3.5 Wang et al. (2025), and MiMo-VL Xiaomi et al. (2025) on its strictly image-disjoint test splits. Figure 1 shows that all four models exhibit substantial degradation across the nine tasks. Adapting each model via SFT on the corresponding training split yields clear improvements, but substantial errors remain, especially for dense and structured outputs. This gap exposes a mismatch between token-level imitation and the requirements of dense visual perception. SFT optimizes the likelihood of a reference response token by token, whereas dense perception is judged at the level of the complete decoded output: whether the correct objects are identified, accurately localized, and reported without omission or redundancy. Token likelihood provides only indirect supervision for these outcome-level requirements and therefore does not directly optimize the semantic, geometric, and structural correctness of the final prediction.

This objective mismatch motivates us to formulate cross-domain adaptation as structured outcome alignment, instantiated through CD-GRPO, a reward-driven post-training framework built on Group Relative Policy Optimization Shao et al. (2024). CD-GRPO decodes each sampled response into a task-specific prediction and assigns sequence-level credit along three complementary dimensions: format validity, task-native accuracy, and structural integrity. The format term enforces parseability, the accuracy term measures semantic and geometric correctness using task-appropriate metrics, and

the structural term evaluates the completeness and consistency of the predicted elements; length-aware control further suppresses redundant and excessive predictions. Unlike conventional visual GRPO formulations centered on compact answer verification, CD-GRPO evaluates dense, variable-length predictions at both the element and set levels. This distinction is critical because an output may be well formatted and partially accurate while still omitting targets, duplicating instances, mismatching categories and locations, or exploiting partial credit through over-prediction. CD-GRPO thus preserves task-specific definitions of correctness within a unified reward framework that operates across heterogeneous tasks and domains without domain-specific redesign.

The experiments support both the diagnosis and the proposed solution. For Qwen3-VL-8B, the overall score averaged across nine tasks and four domains increases from 25.7 without adaptation to 41.7 with SFT and 48.5 with CD-GRPO. CD-GRPO also consistently outperforms three recent perception-oriented RL baselines Yu et al. (2026); Liu et al. (2026); Feng et al. (2026) under equivalent training and evaluation protocols. Reward ablations further isolate the contribution of structure-aware optimization: extending conventional format-and-accuracy GRPO with the complete CD-GRPO formulation improves the overall score from 43.3 to 48.5. Additionally, our scaling analysis examines model sizes from 2B to 8B, target-domain training data at multiple scales, and test-time adaptation through in-context learning and transductive refinement. These three axes characterize how model capacity, target-domain supervision, and test-time computation contribute to robust perception under extreme domain shifts.

Our contributions are threefold: (1) **CrossVLM-Bench**, comprising 7,362 images and 66,258 image–task pairs across four extreme domains, provides a unified testbed for diagnostic evaluation and post-training adaptation across nine perception tasks; (2) **CD-GRPO**, which formulates cross-domain adaptation as structured outcome alignment and directly optimizes complete perceptual predictions; and (3) **Systematic validation**, establishing consistent gains across models and baselines while characterizing adaptation across model capacity, data scale, and test-time strategies.

2 CROSSVLM-BENCH

Data Curation Pipeline. Figure 2 presents the construction pipeline of CrossVLM-Bench. We curate images from four human-annotated source datasets: DIOR for remote sensing Li et al. (2020), NEU-DET for industrial inspection Song & Yan (2013), PlantSeg for precision agriculture Wei et al. (2024), and UODD for underwater imagery Jiang et al. (2021). Because every retained image must support the same nine perception tasks, we select samples with reliable source annotations and sufficient objects, regions, and spatial relationships for multi-task construction. We preserve the original human annotations from each source dataset as ground truth, including category labels and, depending on the dataset, either bounding boxes or segmentation masks. Geometry-dependent targets, such as object counts, are then derived deterministically from these annotations.

GPT-4o Islam & Moushi (2025) assists in converting these source annotations into standardized task-specific instruction–response pairs and generating the associated language content, including referring expressions, reasoning descriptions, and region captions. It does not generate the underlying geometric ground truth. Experienced human annotators subsequently verify and correct every GPT-assisted annotation for visual grounding, semantic consistency, geometry association, format validity, and structural completeness. The resulting benchmark spans these four domains and comprises 7,362 images and 66,258 image–task pairs. Every image supports nine perception tasks: detection, counting, object grounding, referring grounding, semantic segmentation, instance segmentation, referring segmentation, reasoning segmentation, and region captioning. Within each domain, images are partitioned into 60% training and 40% test sets. The split is strictly image-disjoint, and all nine task annotations associated with an image remain in the same partition. CrossVLM-Bench therefore supports both target-domain post-training and held-out evaluation without image-level leakage.

Evaluating Representative VLMs. We first evaluate four representative VLMs on the held-out test split of each domain without target-domain adaptation: Qwen2.5-VL Yang et al. (2025), Qwen3-VL Bai et al. (2025), InternVL3.5 Wang et al. (2025), and MiMo-VL Xiaomi et al. (2025). Figure 1 shows that all four models perform poorly across the nine tasks despite their strong results on standard natural-image benchmarks. The degradation is broad rather than task-specific, affecting object discovery, counting, spatial grounding, segmentation, and region-level understanding across model

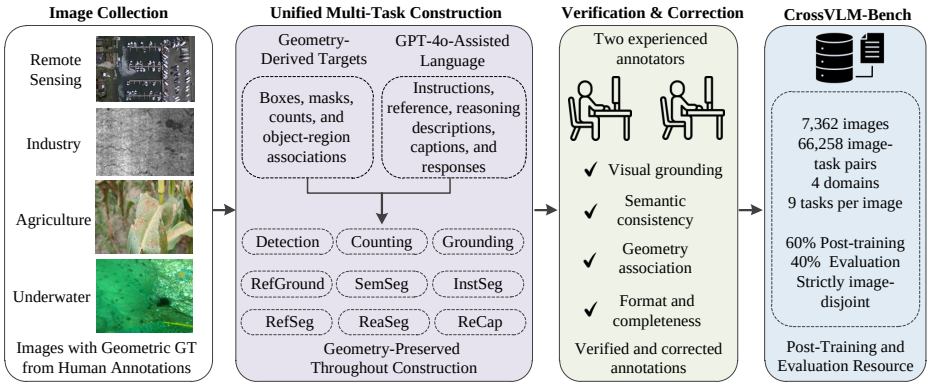

Figure 2: **Construction pipeline of CrossVLM-Bench.** Human-annotated source datasets provide category labels and geometric ground truth in the form of bounding boxes or segmentation masks. GPT-4o assists in transforming these annotations into a unified nine-task format and generating the associated language components, which are subsequently verified and corrected by experienced human annotators. The resulting dataset is divided into strictly image-disjoint training and test splits for post-training and evaluation, respectively.

families. Tasks requiring dense object coverage and precise localization are particularly challenging, while qualitative analysis reveals missed small or crowded objects, repeated or excessive predictions, inaccurate localization, and malformed structured outputs.

These failures are consistent with the substantial differences between web-scale natural imagery and the four evaluated domains in image formation, viewpoint, object scale, texture, illumination, and scene density. They demonstrate that broad task competence on natural images does not by itself ensure robust perception under extreme domain shifts. These results motivate our subsequent study of extreme-domain adaptation, examining how target-domain supervision and reward-driven post-training can restore reliable VLM perception.

3 Adapting VLMs to Extreme Domains

We develop a unified post-training framework for adapting VLMs to extreme visual domains. For each target domain, adaptation uses its training split, while evaluation is performed on the corresponding image-disjoint test split. As illustrated in Fig. 3, SFT first teaches the model the target-domain concepts and structured output interface. CD-GRPO subsequently aligns complete predictions with verifiable perceptual outcomes through group-relative reinforcement learning. Within each domain, the same model handles all nine tasks without task-specific prediction heads.

3.1 Unified Task Formulation

We formulate all nine perception tasks as conditional text generation. Each training sample is represented as $x = (I, q, y^\star, \tau)$, where I is an image, q is a natural-language instruction, $y^\star$ is the reference response, and τ identifies the task type for selecting the corresponding parser and reward during training. The task identifier is not an additional model input because the task is specified by q. For example, an instruction may ask the model to detect all fish or count the apples in an image.

The original annotations are converted into task-specific structured responses containing the required categories, counts, bounding boxes, points, instance identifiers, or captions. For segmentation tasks, the VLM predicts structured spatial prompts, such as boxes and points, which are subsequently converted into pixel masks by SAM Kirillov et al. (2023). The model defines the following autoregressive distribution:

$$\pi_\theta(y^\star \mid I, q) = \prod_{t=1}^{T} \pi_\theta\left(y_t^\star \mid I, q, y_{<t}^\star\right), \tag{1}$$

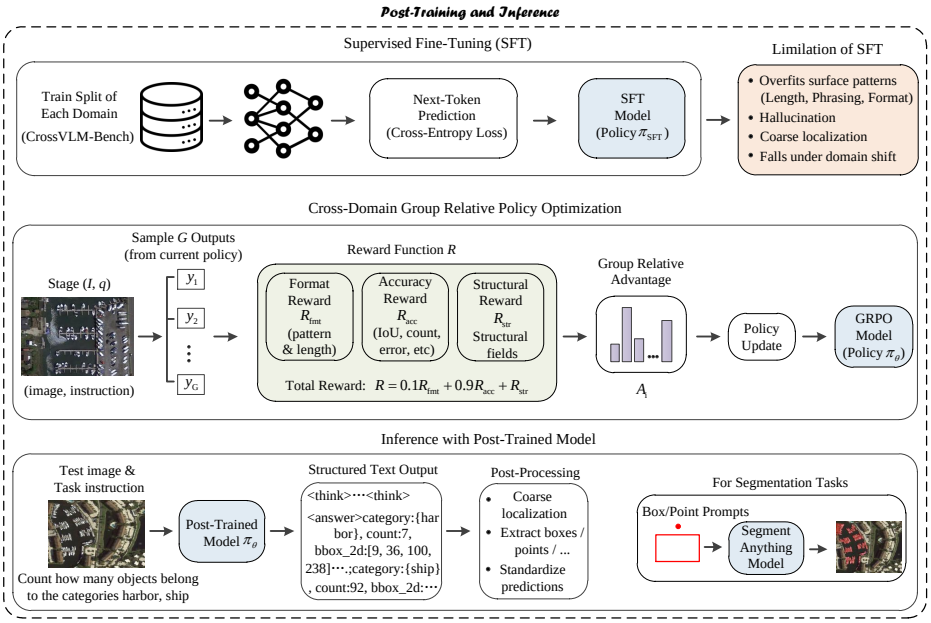

Figure 3: **Post-training and inference framework.** For each target domain, SFT first establishes the unified task interface and initializes the policy. CD-GRPO then optimizes complete predictions using rewards for format validity, task-native accuracy, and structural integrity. At inference, generated responses are parsed into task-specific outputs; for segmentation tasks, predicted boxes and points are passed to a frozen SAM to produce pixel-level masks.

where $y^\star = (y_1^\star, \ldots, y_T^\star)$ is the serialized response and T is its token length. This formulation maps different task outputs into a shared generative interface, allowing one VLM to handle all nine tasks through the same image-instruction-to-response mapping.

3.2 Supervised Fine-Tuning

We first apply SFT to the training split of each target domain. SFT introduces domain-specific visual concepts, teaches the required response schemas, and provides a stable policy initialization for subsequent reinforcement learning. Given the target-domain training set $\mathcal{D}_{\mathrm{SFT}}$, the objective is

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(I,q,y^\star)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\sum_{t=1}^{T}\log \pi_\theta\left(y_t^\star \mid I, q, y_{<t}^\star\right)\right]. \tag{2}$$

SFT substantially improves performance across the target domains, confirming that the training annotations provide effective adaptation signals. Nevertheless, errors remain, particularly for dense and variable-length outputs. The limitation follows from a mismatch between token-level supervision and perceptual evaluation. SFT learns to reproduce one reference response token by token, whereas dense perception is judged from the decoded prediction as a whole, including whether the correct targets are identified, accurately localized, and reported without omissions or redundant predictions. These properties are only indirectly represented by token likelihood. Consequently, a response may follow the expected language and structure while remaining incomplete or geometrically inaccurate. This motivates direct optimization of complete perceptual outcomes.

3.3 Cross-Domain Group Relative Policy Optimization

CD-GRPO formulates extreme-domain adaptation as structured outcome alignment. Built on Group Relative Policy Optimization, it assigns sequence-level rewards to decoded task outputs rather than supervising only the reference token sequence. The reward evaluates whether a response is valid, perceptually accurate, and structurally complete, while controlling redundant predictions. This formulation retains the appropriate correctness criterion for each task within a shared reward framework across tasks and domains.

Group-Relative Policy Update. For each state $s = (I, q)$, the rollout policy $\pi_{\theta_{\text{old}}}$ samples G candidate responses $\{o_i\}_{i=1}^{G}$. After task-specific decoding, each response receives a reward $r_i = R_\tau(o_i, y^\star)$, which is normalized within the group to compute its relative advantage:

$$\widehat{A}_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon}, \qquad \mu_r = \frac{1}{G}\sum_{i=1}^{G} r_i, \qquad \sigma_r = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \mu_r)^2}, \tag{3}$$

where ϵ ensures numerical stability. Responses that perform better than other candidates generated for the same input receive positive advantages, whereas inferior responses receive negative advantages.

The policy is optimized using the clipped GRPO objective

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big(\min\big(\rho_{i,t}\widehat{A}_i,\ \text{clip}(\rho_{i,t}, 1-\delta, 1+\delta)\widehat{A}_i\big) - \beta D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)\Big)\Bigg], \tag{4}$$

where

$$\rho_{i,t} = \frac{\pi_\theta(o_{i,t} \mid s, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid s, o_{i,<t})}. \tag{5}$$

Here, δ controls clipping, and β weights KL regularization. The frozen SFT policy serves as π_{ref}, while $\pi_{\theta_{\text{old}}}$ generates the current rollout group.

Factorized Reward Design. For task τ, CD-GRPO decomposes the reward into format validity, task-native accuracy, and structural integrity:

$$R_\tau = 0.1R_{\text{fmt}} + 0.9R^\tau_{\text{acc}} + R^\tau_{\text{str}}. \tag{6}$$

Format validity. R_{fmt} verifies the outer response structure. It equals 1 when the generated sequence follows the required `<think>...</think><answer>...</answer>` format and 0 otherwise. This term encourages reliable extraction of the final task output from the generated response.

Task-native accuracy. R^τ_{acc} evaluates the semantic or geometric correctness of the decoded prediction using a metric suited to task τ. Detection and object grounding use category-aware localization scores based on IoU. Counting uses normalized absolute count accuracy, while referring grounding uses the IoU of the predicted target box. Referring and reasoning segmentation combine box IoU with point similarity. Instance segmentation uses F1 after matching predicted and reference instances through their boxes and points, whereas semantic segmentation uses greedy region matching with the same spatial signals. Region captioning is evaluated using semantic similarity. For tasks producing multiple records, the accuracy reward includes a length-aware penalty that suppresses excessive predictions.

Structural integrity. $R^\tau_{\text{str}} \in [0, 0.5]$ evaluates whether the decoded output is complete and internally consistent. For detection and object grounding, it measures coverage of the required object categories. For counting, it checks consistency between predicted counts and the corresponding numbers of boxes. Referring grounding requires exactly one target box, while referring and reasoning segmentation require one box-point pair. For instance and semantic segmentation, the reward measures the fraction of valid instance or region records and penalizes severe over-prediction. The task metadata associated with each training sample selects the corresponding accuracy and structural criteria automatically.

Together, these components move reward optimization beyond format compliance and partial task accuracy. CD-GRPO evaluates both individual predictions and the complete output structure, directly discouraging missing targets, duplicated instances, inconsistent fields, and excessive predictions.

3.4 Inference and Task Decoding

At inference, the adapted VLM generates a structured response conditioned only on the image and instruction. A deterministic parser extracts the content enclosed by the answer tags, validates the task-specific schema, and converts normalized textual predictions into their corresponding task

outputs. Detection, counting, grounding, and region-captioning results are evaluated directly from the parsed response. For segmentation tasks, the predicted boxes and points are provided to a frozen SAM to generate pixel-level masks. The same SAM checkpoint and prompting protocol are used for every evaluated VLM and adaptation method, ensuring that performance differences arise from the spatial prompts generated by the VLM rather than changes to the mask decoder.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data. All experiments use CrossVLM-Bench, spanning remote sensing, industrial inspection, agriculture, and underwater imagery. Each domain is partitioned at the image level into 60% training and 40% test data for post-training and evaluation, respectively. For CD-GRPO, we retain training samples whose SFT accuracy lies within an intermediate range around 0.5, focusing optimization on informative examples that are neither trivial nor excessively difficult. The selected samples are manually checked to remove ambiguous or inconsistent cases.

Models. We evaluate four representative VLMs as baselines: Qwen-VL-2.5 (7B) Yang et al. (2025), Qwen-VL-3.0 (8B) Bai et al. (2025), MiMo-VL (7B) Xiaomi et al. (2025), and InternVL-3.5 (8B) Wang et al. (2025). For each model, we report direct inference performance on the test split (baseline), as well as performance after post-training (SFT and CD-GRPO). Each model is adapted independently to each target domain using the corresponding training split.

Implementation Details. For both SFT and CD-GRPO, we use AdamW with a learning rate of 3×10^{-5} and cosine learning-rate decay. SFT uses a batch size of 8, while CD-GRPO uses a group size of $G = 5$, a clipping coefficient of $\delta = 10^{-2}$, and a KL coefficient of $\beta = 0.01$. The trainable policy is initialized from the SFT checkpoint, and a frozen copy of the same checkpoint serves as the reference policy. SFT and CD-GRPO are each trained for one epoch. Unless otherwise specified, all domains share the same hyperparameters. All experiments run on eight NVIDIA H100 GPUs.

Evaluation Metrics. We evaluate each task using its corresponding outcome-level metric. Detection is measured by average precision (AP). Counting accuracy is the percentage of examples for which the predicted count exactly matches the ground truth. Object grounding and referring grounding are evaluated by the proportion of predictions whose intersection over union satisfies $\mathrm{IoU} > 0.5$. Semantic, instance, referring, and reasoning segmentation are evaluated using mean intersection over union (mIoU); the spatial prompts predicted by the VLM are converted into pixel-level masks using the same frozen SAM checkpoint and prompting procedure for all methods. Region captioning is assigned a binary correctness score by a fixed LLM judge that evaluates semantic correctness and visual grounding under the same prompt for every method.

4.2 MAIN RESULTS

Fig. 1 summarizes the domain-level score for each model, computed as the unweighted mean across the nine tasks. Across all evaluated VLMs and domains, direct inference without target-domain adaptation yields poor performance, SFT provides substantial recovery, and CD-GRPO consistently achieves the strongest results. For Qwen3-VL-8B, CD-GRPO outperforms the baseline by 14.8, 28.0, 14.4, and 33.9 points on remote sensing, industrial inspection, agriculture, and underwater imagery, respectively, while exceeding SFT by a further 6.6, 3.4, 6.4, and 10.6 points. CD-GRPO likewise outperforms SFT for Qwen2.5-VL, InternVL3.5, and MiMo-VL across all four domains, demonstrating consistent effectiveness across distinct VLM families. Complete task-wise results in the appendix further show that these gains span semantic, geometric, and language-grounded perception rather than being concentrated in a single task.

4.3 COMPARISON WITH PERCEPTION-ORIENTED RL

We compare CD-GRPO with three recent perception-oriented RL methods: Perception-R1 Yu et al. (2026), VisionReasoner Liu et al. (2026), and OneThinker Feng et al. (2026). To ensure a controlled comparison, we reproduce all methods using the same Qwen3-VL-8B backbone, target-domain training images, nine-task output space, and test splits. As shown in Table 1, all RL

Table 1: **Comparison with perception-oriented RL methods.** All methods use Qwen3-VL-8B and the same training and evaluation protocol.

Method	DIOR	NEU-DET	PlantSeg	UODD
Baseline	26.8	28.2	19.3	28.4
Perception-R1 Yu et al. (2026)	34.2	41.5	25.4	48.9
VisionReasoner Liu et al. (2026)	37.1	44.8	28.6	51.7
OneThinker Feng et al. (2026)	38.6	47.6	25.9	50.9
CD-GRPO (ours)	**41.6**	**56.2**	**33.7**	**62.3**

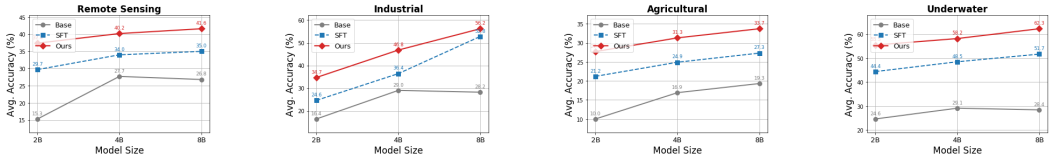

Figure 4: **Model-capacity scaling across domains.**

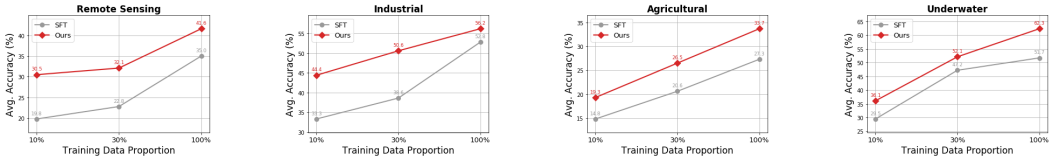

Figure 5: **Data efficiency across domains.**

methods improve substantially over base model, confirming the general value of outcome-driven post-training. CD-GRPO nevertheless achieves the best result in every domain, exceeding the strongest competing method by 3.0–10.6 points.

4.4 Ablation Studies

Training Stages. We first examine whether the improvement arises merely from longer optimization or additional training computation. Table 2(a) compares SFT, extended SFT, direct GRPO without SFT initialization, and the complete two-stage recipe. Extending SFT to the same two-epoch budget consistently degrades performance, ruling out training duration as the source of the gain. Direct GRPO improves over the base model but remains clearly below the complete method. These results reveal complementary roles for the two stages: SFT establishes the target-domain task and output interface, while CD-GRPO aligns complete predictions with verifiable outcomes.

Reward Design. Table 2(b) disentangles the roles of the three reward components. The format-only variant enforces a valid response schema but offers little supervision for semantic or geometric correctness. Incorporating task-native accuracy produces the dominant improvement and recovers the format-and-accuracy objective commonly used in visual GRPO. Structural supervision nevertheless contributes a further 3.8–6.5 points across all four domains by accounting for omissions, redundancy, and internal inconsistencies that accuracy scores alone do not fully capture. The complete objective performs best by jointly optimizing prediction accuracy and output integrity.

Table 2: **Ablations of CD-GRPO.** In (a), 1+1 denotes one SFT epoch followed by one CD-GRPO epoch. In (b), R_{fmt}, R_{acc}, and R_{str} denote format validity, task-native accuracy, and structural integrity, respectively.

(a) Training-stage ablation

Method	Epochs	DIOR	NEU-DET	PlantSeg	UODD
Baseline	–	26.8	28.2	19.3	28.4
SFT	1	35.0	52.8	27.3	51.7
SFT-Long	2	33.6	50.9	25.8	49.4
GRPO w/o SFT	2	35.9	53.7	28.6	52.8
CD-GRPO (ours)	1+1	**41.6**	**56.2**	**33.7**	**62.3**

(b) Reward-component ablation

Reward	DIOR	NEU-DET	PlantSeg	UODD
R_{fmt}	12.4	17.8	10.0	19.2
$R_{\text{fmt}} + R_{\text{str}}$	17.1	23.6	13.7	25.9
$0.1R_{\text{fmt}} + 0.9R_{\text{acc}}$	37.2	50.4	29.9	55.8
CD-GRPO (ours)	**41.6**	**56.2**	**33.7**	**62.3**

4.5 Scalability Analysis

Model Capacity. Fig. 4 reveals a clear distinction between model scaling and adaptation scaling. Increasing the size of an unadapted VLM does not consistently improve extreme-domain perception, indicating that additional capacity alone cannot overcome the distribution shift. Once adapted, however, larger models benefit more reliably, and CD-GRPO consistently outperforms SFT across model sizes and domains.

Data Efficiency. Fig. 5 compares SFT and CD-GRPO across different target-domain data budgets. Performance improves consistently as more supervision becomes available, confirming that extreme-domain adaptation continues to benefit from additional data. CD-GRPO outperforms SFT at every data budget and in every domain, including the low-data regime, showing that its effectiveness does not depend on abundant annotations. Its advantage also persists as the training set grows, indicating that outcome-level optimization complements data scaling. These results show that both supervision scale and adaptation objective are important: additional data expands target-domain knowledge, while CD-GRPO exploits that supervision more effectively.

Test-Time Adaptation. We further examine whether target-domain context can refine the post-trained model at inference time. In-context learning conditions each query on a small set of labeled demonstrations from the corresponding domain, providing task and domain guidance through the prompt. Transductive refinement instead aggregates the model's previous predictions across tasks for the same query image, restores their spatial and semantic context, and prompts the model to re-examine and correct its outputs in a subsequent inference round. As shown in Table 3, both consistently improve CD-GRPO without parameter updates, offering lightweight refinement when retraining is impractical,

Table 3: **Test-time adaptation.** Both ICL and transductive refinement improve CD-GRPO without parameter updates.

Method	Remote	Industrial	Agricultural	Underwater
CD-GRPO	41.6	56.2	33.7	62.3
+ICL	**43.8**	57.4	35.1	63.9
+Trans.	43.3	**58.3**	**35.4**	**64.0**

5 Related Work

Generalist VLMs and Cross-Domain Learning. Modern VLMs such as Qwen3-VL Bai et al. (2025), InternVL3.5 Wang et al. (2025), and MiMo-VL Xiaomi et al. (2025) unify diverse perception and reasoning tasks through a shared generative interface. Existing benchmarks evaluate grounding, segmentation, and multimodal understanding Yu et al. (2016); Lai et al. (2024); Liu et al. (2024a); Li et al. (2024), but remain dominated by natural imagery; they show task breadth without testing robustness to shifts in imaging processes and visual statistics. CrossVLM-Bench fills this gap by supporting post-training and evaluation across extreme visual domains. Cross-domain learning has addressed distribution shift via domain generalization, adaptation, and few-shot learning Guo et al. (2020); Zhou et al. (2023); Lei et al. (2022); Venkateswara et al. (2017); Peng et al. (2019); Choe et al. (2025); Zhu et al. (2025), yet prior work typically handles classification, segmentation, or detection separately with task-specific architectures and objectives Chen et al. (2024); Fu et al. (2025); Dai et al.. In contrast, we study whether a generative VLM can preserve nine semantic, geometric, and language-grounded capabilities through a common interface under domain shift.

Reinforcement Learning for Visual Perception. GRPO Shao et al. (2024) has recently emerged as an effective framework for reinforcement learning with verifiable visual outcomes. Perception-R1 Yu et al. (2026) examines how task characteristics and rule-based rewards affect perception learning; VisionReasoner Liu et al. (2026) integrates task reformulation and multi-object reasoning within a unified perception framework; and OneThinker Feng et al. (2026) introduces EMA-GRPO to stabilize optimization across heterogeneous image and video tasks. Collectively, these methods establish the effectiveness of reinforcement learning for improving visual perception and reasoning.

CD-GRPO extends this direction to a complementary setting and differs along three dimensions. (1) Its objective is extreme-domain adaptation rather than capability enhancement on standard visual benchmarks, requiring a VLM to support nine semantic, geometric, and language-grounded tasks through a common interface after substantial distribution shift. (2) Its unit of verification is the complete decoded prediction rather than an individual answer, reasoning trajectory, or aggregate task score. This distinction is essential for dense, variable-length outputs, where locally accurate elements may coexist with missing targets, duplicate instances, inconsistent semantic-spatial associations, or excessive predictions. (3) Its reward formulation separates task-native fidelity from task-shared structural constraints: the former preserves the appropriate semantic or geometric metric for each task, while the latter enforces output validity, completeness, consistency, and controlled cardinality. The resulting factorization is shared across all nine tasks and retained when adapting independently to each extreme domain. CD-GRPO therefore advances visual reinforcement learning from task-level performance optimization toward structured outcome alignment under severe domain shift.

6 Conclusion

This work identifies robustness to extreme visual domains as an underexplored dimension of VLM generality. CrossVLM-Bench exposes systematic degradation across four domains and nine tasks, while revealing the limitations of adapting VLMs through token-level imitation alone. CD-GRPO addresses this objective mismatch by optimizing decoded predictions as complete structured outcomes, consistently outperforming SFT and recent perception-oriented RL methods. Our scaling analyses further clarify how model capacity, target-domain data, and test-time adaptation shape adaptation effectiveness. Together, these contributions provide a systematic foundation for evaluating and improving VLMs under extreme domain shifts.

AI USE STATEMENT

In this work, we used generative AI tools for two purposes. First, they assisted with manuscript preparation, including translation, language polishing, and refinement of wording, sentence structure, and presentation. Second, GPT-4o was used during the construction of CrossVLM-Bench to generate initial candidate task and language annotations from the human-annotated geometric ground truth of the source datasets. We did not use generative AI to formulate mathematical claims or proofs, and the remaining proof-related disclosure tasks are not applicable to this work.

All AI-assisted text was reviewed and revised by the authors. Every AI-generated candidate annotation was manually verified and corrected before inclusion in the benchmark; the source images and geometric ground truth were not generated by AI. We take full responsibility for the final content of this work, including its text, annotations, claims, results, and artifacts.

ETHICS STATEMENT

This work uses only publicly available datasets collected under their original licenses; no human-subject experiments were conducted. Generalist VLMs adapted to extreme visual domains may inherit biases from pretraining data and downstream benchmarks, leading to uneven performance across domains or sensors. Our reward factorization improves output validity and consistency but does not by itself guarantee fairness. We declare no conflicts of interest or sponsorship influencing the findings.

REPRODUCIBILITY STATEMENT

We provide the model backbone, post-training procedure, reward factorization, and test-time adaptation strategies in the main text. Hyperparameters, data splits, reward definitions, task-specific processing, and the domain adaptation protocol are detailed in the Appendix and supplementary materials. Source code, reward implementations, and evaluation scripts will be released upon acceptance.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Abderrahmene Boudiaf, Irfan Hussain, and Sajid Javed. Agrichat: A multimodal large language model for agriculture image understanding. *arXiv preprint arXiv:2603.16934*, 2026.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, and Jungong Han. Pixel matching network for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 978–987, 2024.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Seun-An Choe, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Universal domain adaptation for semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4607–4617, 2025.

Boyang Dai, Zeng Fan, Zihao Qi, Meng Lou, and Yizhou Yu. Cgsa: Class-guided slot-aware adaptation for source-free object detection. In *The Fourteenth International Conference on Learning Representations*.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, 2026.

Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qi, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. *COMPUTER VISION-ECCV 2024, PT LVIII*, 15116:247–264, 2025.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pp. 124–141. Springer, 2020.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 47–60. Springer, 2025.

Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4259–4267, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9579–9589, 2024.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European conference on computer vision*, pp. 73–90. Springer, 2022.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

Qingyun Li, Shuran Ma, Junwei Luo, Yi Yu, Yue Zhou, Fengxiang Wang, Xudong Lu, Xiaoxing Wang, Xin He, Yushi Chen, et al. Co-training vision-language models for remote sensing multi-task learning. *Remote Sensing*, 18(2):222, 2026a.

Yanhui Li, Yunkang Cao, Chengliang Liu, Yuan Xiong, Xinghui Dong, and Chao Huang. Iad-r1: Reinforcing consistent reasoning in industrial anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6583–6591, 2026b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3170–3180, 2023.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Matvei Popov, Peter Robicheaux, Anish Madan, Isaac Robinson, Joseph Nelson, Deva Ramanan, and Neehar Peri. Roboflow100-vl: A multi-domain object detection benchmark for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv e-prints*, pp. arXiv–2402, 2024.

Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285:858–864, 2013. ISSN 0169-4332. doi: https://doi.org/10.1016/j.apsusc.2013.09.002. URL https://www.sciencedirect.com/science/article/pii/S0169433213016437.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Tianqi Wei, Zhi Chen, Xin Yu, Scott Chapman, Paul Melloy, and Zi Huang. Plantseg: A large-scale in-the-wild dataset for plant disease segmentation. *arXiv preprint arXiv:2409.04038*, 2024.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025.

Wei Xu, Cheng Wang, Dingkang Liang, Zongchuang Zhao, Xingyu Jiang, Peng Zhang, and Xiang Bai. Nautilus: A large multimodal model for underwater scene understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2. 5-1m technical report. 2025.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:94827–94853, 2026.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20061–20070, 2023.

Ronghang Zhu, Mengxuan Hu, Weiming Zhuang, Lingjuan Lyu, Xiang Yu, and Sheng Li. Revisiting source-free domain adaptation: Insights into representativeness, generalization, and variety. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25688–25697, 2025.

Appendix Organization. This appendix provides supplementary documentation and experiments supporting the main paper. Section A details the data sources, curation criteria, annotation pipeline, task definitions, and quality-control procedures of CrossVLM-Bench. Section B reports the complete task-level results across all models and domains. Section C describes the implementation protocols and practical implications of the model-capacity, data-efficiency, and test-time adaptation studies.

Section D.1, Section D.2, and Section D.3 provide controlled comparisons with task-specialist models, domain-specialist VLMs, and a closed-source VLM, respectively. Section E.1, Section E.2, Section E.3, and Section E.4 examine general-capability retention, run-to-run stability, the contributions of VLM localization and SAM, and representative failure modes. Finally, Section F presents qualitative comparisons, and Section G discusses limitations and future directions.

A CrossVLM-Bench Details

A.1 Data Collection and Domain Characteristics

CrossVLM-Bench covers four extreme visual domains: remote sensing, industrial inspection, agriculture, and underwater imagery. We construct each domain from a publicly available source dataset with human-annotated geometric supervision. The available category labels, bounding boxes, and/or segmentation masks are preserved as ground truth rather than regenerated by GPT-4o. Geometry-dependent targets, including object counts and object–region associations, are derived deterministically from these annotations.

Our subset construction is task-driven rather than a random reduction of the source datasets. Because every retained image must meaningfully support the same nine tasks, we select images with reliable source geometry and sufficient objects, regions, and spatial relations for constructing detection, segmentation, counting, grounding, referring, reasoning, and captioning examples. Selection is independent of model predictions and preserves the complete category vocabulary of each source dataset. We additionally retain challenging cases involving small or densely distributed objects, occlusion, irregular geometry, complex layouts, heterogeneous backgrounds, and domain-specific image degradation.

- **Remote Sensing.** We use DIOR Li et al. (2020), which contains aerial scenes covering 20 categories, including airplanes, ships, vehicles, harbors, and storage tanks. The selected 2,100 images emphasize small objects, dense spatial distributions, multi-category scenes, and complex geographic layouts, with approximately 40 instances per image.
- **Industrial Inspection.** We use NEU-DET Song & Yan (2013), which contains six types of steel surface defects: crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches. Its 1,800 images contain irregular defect geometry, substantial texture variation, and approximately 23 annotated instances per image.
- **Agriculture.** We use PlantSeg Wei et al. (2024), an in-the-wild plant-disease dataset spanning 115 categories. We retain 2,100 images containing diverse plant species and symptoms, overlapping leaves, scale variation, and heterogeneous natural backgrounds, with approximately 16 instances per image.
- **Underwater Imagery.** We use UODD Jiang et al. (2021), which contains sea cucumbers, sea urchins, and scallops. Its 1,362 selected images exhibit occlusion, color distortion, turbidity, nonuniform illumination, and approximately 19 instances per image.

Table 4 summarizes the resulting benchmark. CrossVLM-Bench contains 7,362 images and 66,258 image–task pairs. Within each domain, images are divided into approximately 60% training and 40% test data using a strictly image-disjoint split. All nine annotations derived from one image remain in the same partition, so no annotation associated with a test image is used during SFT or CD-GRPO training.

A.2 GPT-4o-Assisted Task Construction

Fig. 2 illustrates the complete construction pipeline. The benchmark combines trusted source geometry, deterministic target derivation, GPT-4o-assisted language construction, and human verification. The pipeline comprises four stages.

Table 4: **Domain statistics of CrossVLM-Bench.** Each image supports all nine perception tasks.

Domain	Images	Classes	Inst./image
Remote sensing	2,100	20	40
Industrial	1,800	6	23
Agricultural	2,100	115	16
Underwater	1,362	3	19
Total	**7,362**	–	–

Source-Annotation Preservation. We first normalize the human-annotated category labels and available geometric annotations from each source dataset into a common representation. The original boxes and/or masks remain the authoritative geometric ground truth throughout benchmark construction.

Geometry-Derived Targets. Object counts, category–instance correspondences, object–region associations, and task-specific geometric targets are derived deterministically from the normalized source annotations. This prevents GPT-4o hallucinations from altering the underlying detection or segmentation supervision.

GPT-4o-Assisted Task and Language Construction. Conditioned on the image, source annotations, domain information, and task specification, CoT-prompted GPT-4o Islam & Moushi (2025) constructs task instructions, referring expressions, reasoning descriptions, region captions, and structured responses. The prompt requires the model to ground every generated statement in the supplied image and source geometry, maintain correspondence between textual references and object instances, and follow the output schema of the corresponding task. We perform up to three refinement rounds in which GPT-4o checks its draft for missing references, unsupported attributes, inconsistent counts, and malformed fields. Fig. 6 presents an example prompt template.

Human Verification. The complete benchmark, comprising all 66,258 image–task pairs, is verified by two laboratory members with prior annotation experience. They examine visual grounding, semantic correctness, task completeness, format validity, and correspondence with the source geometry. Residual errors in referring expressions, captions, reasoning descriptions, counts, and structured fields are corrected manually. Importantly, this stage verifies the GPT-4o-assisted language and task construction; it does not replace the source datasets' human-annotated geometric ground truth.

A.3 Task Definitions

Every image is associated with the same nine task types, yielding complementary supervision over recognition, localization, segmentation, language grounding, and region-level understanding.

- **Detection (Det):** predict the class label and bounding box of each target object.
- **Counting (Cnt):** predict the number of instances belonging to the queried category.
- **Object Grounding (Grd):** localize all objects matching a category or natural-language query.
- **Referring Grounding (RefG):** localize the specific instance described by a referring expression.
- **Semantic Segmentation (Seg):** identify the complete pixel region associated with a queried semantic category.
- **Instance Segmentation (ISeg):** distinguish and segment individual object instances of the target category.
- **Referring Segmentation (RefS):** segment the object or region specified by a referring expression.
- **Reasoning Segmentation (ReaS):** infer and segment the target region from a query requiring semantic or relational reasoning.

```
messages = ["role": "system", "content": "You are generating a visual reasoning chain for a SEMANTIC SEGMENTATION task on the industrial domain.
CONTEXT:
You will be given a QUESTION and a set of KNOWN VISUAL FINDINGS.
The KNOWN VISUAL FINDINGS do NOT represent labels or ground truth; instead, they reflect visual details that are progressively discovered and confirmed during image inspection.
YOUR GOAL:
Write a realistic, step-by-step visual reasoning process that naturally leads to those findings, as if inferred directly from the image while performing semantic segmentation for the requested categories.
The category space is defects: crazing, inclusion, patches, pitted_surface, rolled-in_scale, scratches.
OUTPUT CONSTRAINTS:
- Output ONLY the reasoning text itself.
- Do NOT include <think>, <answer>, markdown, numbering, or any other formatting.
- Do NOT explicitly mention terms such as 'given answer', 'provided findings', 'provided bbox', 'coordinates are given', 'annotation', 'label', or 'ground truth'.
- Do NOT invent objects, categories, or any new coordinates.
- Stay in industrial steel-surface context; do not introduce natural-scene objects (cars, roads, buildings, sky, trees, etc.).
- Do NOT assign a specific class name to a bbox unless it is directly supported by the QUESTION and clearly visible; otherwise refer to it as a target instance belonging to one of the requested categories.
SPATIAL DESCRIPTION GUIDELINES:
- If coordinates appear in the reasoning, use EXACTLY these formats: bbox_2d:[x1, y1, x2, y2] and point_2d:[x, y]
- The numbers must match the KNOWN VISUAL FINDINGS exactly.
- Present coordinates as the result of visual inspection and refinement, not as externally provided information.
REASONING STYLE:
- Begin with a global scan of the steel surface and interpret the segmentation goal from the QUESTION.
- Gradually narrow attention to task-relevant regions.
- For each bbox_2d/point_2d pair, describe how you identify the region, use point_2d as an interior seed/anchor, and refine the defect boundary (edges, shape, texture, contrast, continuity breaks) to include defect pixels and exclude normal background.
- If there are many instances, you may group similar ones by spatial clusters, but you MUST reference every bbox_2d/point_2d pair at least once.
- Conclude with a brief verification that all requested-category regions have been covered."]
for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample["context"]})
    messages.append({"role": "assistant", "content": sample["response"]})

messages.append({"role":"user", "content":'\n'.join(query)})
```

Figure 6: **Example task-construction prompt.** The template conditions GPT-4o on the image, task specification, and source geometry to construct grounded language and structured task outputs. Core geometric supervision is inherited from the human-annotated source dataset.

- **Region Captioning (ReCap):** generate a visually grounded description of a specified image region.

Fig. 7 shows representative examples from the four domains and illustrates how the nine tasks provide complementary views of the same visual evidence. The resulting annotations support both target-domain post-training on the training split and unified evaluation on the held-out test split.

A.4 Independent Annotation-Quality Audit

Beyond the primary verification process, we conduct an independent, stratified audit of 120 images, with 30 images sampled from each domain and coverage of both training and test partitions. Because all nine task annotations are reviewed for every image, the audit covers 1,080 image–task pairs. Two annotators independently inspect each original GPT-4o-assisted output, and disagreements are resolved by a third annotator.

An output is marked *Accept* when its language and structured task components require no correction, and *Revise* when human intervention is required to repair an unsupported attribute, ungrounded description, answer-first reasoning pattern, incorrect association, or structural error. As shown in Table 5, 81.6% of the GPT-4o-assisted outputs are accepted directly, while 18.4% require revision. The lower acceptance rate in agriculture mainly reflects the difficulty of grounding descriptions in cluttered scenes with overlapping leaves and subtle disease symptoms.

These errors arise from language grounding and task-format construction, rather than from the underlying boxes or masks inherited from the source datasets. The audit therefore separates the reliability of geometric supervision from the quality of GPT-4o-assisted language components.

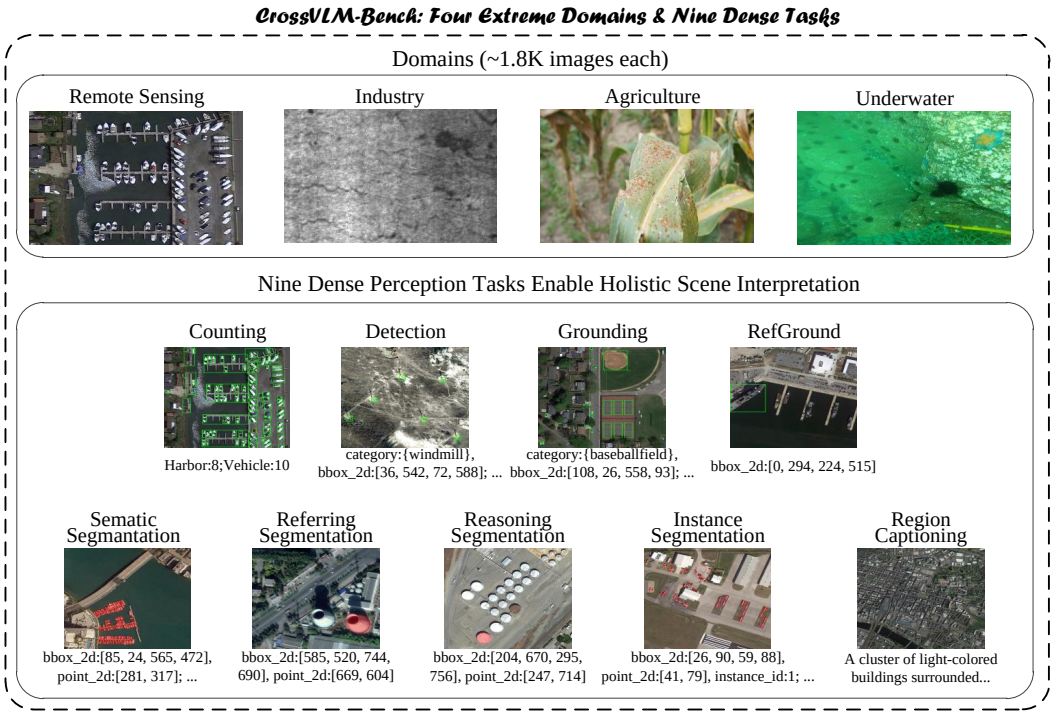

Figure 7: **CrossVLM-Bench task examples.** Representative examples from remote sensing, industrial inspection, agriculture, and underwater imagery, covering all nine perception tasks.

Table 5: **Independent audit of GPT-4o-assisted annotations.** Each domain contains 30 audited images and 270 task annotations.

Domain	Accept (%)	Revise (%)	Most common error
Remote sensing	83.6	16.4	Answer-first reasoning
Industrial	89.2	10.8	Hallucinated attribute
Agricultural	74.5	25.5	Ungrounded description
Underwater	78.9	21.1	Hallucinated attribute
Average	**81.6**	**18.4**	Answer-first reasoning

A.5 Annotation Efficiency

We additionally compare the adopted pipeline with annotation performed entirely by humans. On the audited subset, constructing all nine task annotations manually requires 14.2 minutes per image on average, whereas GPT-4o-assisted construction followed by human verification requires 4.8 minutes. The pipeline therefore reduces annotation time by 66.2% while retaining full human verification.

A.6 Comparison with Existing Benchmarks

Table 7 compares CrossVLM-Bench with three related benchmark families.

Natural-Image Multimodal Benchmarks. RefCOCO Yu et al. (2016), ReasonSeg Lai et al. (2024), and CountBench Paiss et al. (2023) evaluate important VLM capabilities, including referring localization, reasoning segmentation, and counting. However, they are centered on natural imagery and individual task families, providing limited evidence about whether a unified VLM retains these capabilities under substantial changes in imaging conditions and visual statistics.

Cross-Domain Learning Benchmarks. Meta-Dataset Triantafillou et al., CD-FSL Guo et al. (2020), CD-FSS Lei et al. (2022), CD-FSOD Fu et al. (2025), and Roboflow100-VL Popov et al. (2025) explicitly study distribution shift. Nevertheless, each benchmark primarily addresses one prediction problem, such as classification, segmentation, or detection. They therefore cannot determine

Table 6: **Annotation efficiency.** Time denotes the average required to construct and verify all nine task annotations for one image.

Pipeline	Min./image	Time reduction
Fully manual	14.2	–
GPT-4o assisted + verification	4.8	66.2%

Table 7: **Comparison with representative benchmarks.** CrossVLM-Bench combines extreme-domain shift with dense multi-task annotations for VLM post-training and evaluation.

Category	Benchmark	Cross-domain	Task scope
Multimodal	RefCOCO Yu et al. (2016)	No	Referring grounding/segmentation
Multimodal	ReasonSeg Lai et al. (2024)	No	Reasoning segmentation
Multimodal	CountBench Paiss et al. (2023)	No	Counting
Cross-domain	Meta-Dataset Triantafillou et al.	Yes	Classification
Cross-domain	CD-FSL Guo et al. (2020)	Yes	Classification
Cross-domain	CD-FSS Lei et al. (2022)	Yes	Segmentation
Cross-domain	CD-FSOD Fu et al. (2025)	Yes	Detection
Cross-domain	Roboflow100-VL Popov et al. (2025)	Yes	Detection
Adaptation	Office-Home Venkateswara et al. (2017)	Yes	Classification
Adaptation	DomainNet Peng et al. (2019)	Yes	Classification
Ours	**CrossVLM-Bench**	**Yes**	**Nine perception tasks**

whether multiple semantic, geometric, and language-grounded abilities transfer together through a shared generative interface.

Domain-Adaptation Benchmarks. Office-Home Venkateswara et al. (2017) and DomainNet Peng et al. (2019) provide standard settings for studying domain adaptation, but are dominated by image classification and do not offer dense, per-image supervision for adapting and assessing the broad perceptual capabilities of a VLM.

Distinguishing Scope. CrossVLM-Bench is designed to connect two functions that are usually separated: diagnosing the breakdown of unified VLM perception under extreme domain shift and providing target-domain supervision for post-training adaptation. Its defining property is not merely the number of tasks or domains, but the alignment of nine complementary task annotations on every image under strictly image-disjoint training and test splits. This design makes it possible to study domain robustness as a unified model capability while retaining task-level diagnostic resolution.

Across all sixteen model–domain combinations, CD-GRPO improves upon SFT. The gains extend across semantic, geometric, and language-grounded tasks, supporting the architecture- and domain-level consistency reported in the main paper.

B Complete Task-Level Results

The main paper reports domain-level means across nine tasks. Here, we provide complete task-level results for all four VLM families. All models are adapted independently on the corresponding training split and evaluated on the image-disjoint test split. Scores are reported on a 0–100 scale, and Avg. is the unweighted mean over nine tasks.

As shown in Tables 8, 9, 10, and 11, CD-GRPO outperforms SFT across all sixteen model–domain combinations. The improvements span semantic, geometric, and language-grounded tasks, corroborating the architecture- and domain-level consistency summarized in the main paper.

Table 8: Performance on the remote sensing domain of CrossVLM-Bench.

Model	Method	Task									
		Det	Count	Ground	Ref-Ground	Seg	InstSeg	RefSeg	ReaSeg	ReCap	Avg
Qwen2.5-VL-7B Yang et al. (2025)	Baseline	4.2	15.4	6.1	9.2	18.9	20.2	40.3	32.6	9.8	17.4
	+SFT	13.1	25.1	15.7	34.7	26.5	25.7	49.8	40.8	26.5	28.7
	+CD-GRPO	**16.5**	**32.4**	**20.3**	**36.9**	**34.7**	**31.4**	**55.3**	**45.7**	**30.8**	**33.8**
Qwen3-VL-8B Bai et al. (2025)	Baseline	18.8	23.7	16.1	40.9	16.1	17.5	44.4	34.3	29.1	26.8
	+SFT	25.3	27.5	28.7	43.8	28.7	29.1	52.5	48.0	31.3	35.0
	+CD-GRPO	**31.4**	**33.9**	**37.2**	**45.1**	**44.8**	**35.9**	**60.5**	**51.4**	**34.5**	**41.6**
MiMo-VL-7B Xiaomi et al. (2025)	Baseline	9.7	24.0	5.3	25.5	10.6	10.4	23.3	26.4	11.2	16.3
	+SFT	12.6	25.3	16.8	36.8	18.5	15.7	30.2	29.8	18.6	22.7
	+CD-GRPO	**13.6**	**29.1**	**22.9**	**37.6**	**21.7**	**26.1**	**40.9**	**36.3**	**24.6**	**28.1**
InternVL3.5-8B Wang et al. (2025)	Baseline	24.3	42.1	19.1	38.2	16.7	15.9	31.1	22.2	13.6	24.8
	+SFT	26.6	42.8	26.9	38.7	25.3	16.5	39.8	40.1	30.1	31.9
	+CD-GRPO	**28.2**	**44.2**	**34.2**	**40.2**	**27.5**	**29.7**	**47.2**	**48.3**	**36.6**	**37.3**

Table 9: Performance on the industrial domain of CrossVLM-Bench.

Model	Method	Task									
		Det	Count	Ground	Ref-Ground	Seg	InstSeg	RefSeg	ReaSeg	ReCap	Avg
Qwen2.5-VL-7B Yang et al. (2025)	Baseline	6.4	24.0	1.1	13.7	12.8	10.1	17.1	14.8	12.6	12.5
	+SFT	22.2	43.2	24.9	18.5	13.6	25.2	28.5	24.0	23.8	24.9
	+CD-GRPO	**24.6**	**46.8**	**36.9**	**31.0**	**38.8**	**49.1**	**56.6**	**56.5**	**35.9**	**41.8**
Qwen3-VL-8B Bai et al. (2025)	Baseline	12.2	25.9	28.6	39.5	30.5	16.2	39.9	35.3	25.5	28.2
	+SFT	51.0	52.4	56.6	57.5	42.4	47.8	66.6	65.4	35.8	52.8
	+CD-GRPO	**52.1**	**54.6**	**58.8**	**58.2**	**48.8**	**50.1**	**71.2**	**69.7**	**42.3**	**56.2**
MiMo-VL-7B Xiaomi et al. (2025)	Baseline	5.7	23.5	6.8	13.2	11.0	6.5	9.0	11.3	13.8	11.2
	+SFT	21.5	36.3	23.4	15.6	13.1	10.8	14.4	13.6	25.9	19.4
	+CD-GRPO	**29.5**	**47.7**	**41.3**	**31.9**	**33.7**	**43.7**	**53.7**	**51.3**	**38.9**	**41.3**
InternVL3.5-8B Wang et al. (2025)	Baseline	5.3	26.9	6.3	15.5	11.4	4.4	5.4	12.9	11.3	11.0
	+SFT	56.0	51.0	58.9	60.7	52.6	52.7	68.8	62.2	21.8	53.9
	+CD-GRPO	**59.2**	**53.1**	**61.5**	**62.3**	**54.8**	**56.8**	**72.1**	**65.8**	**29.8**	**57.3**

Table 10: Performance on the agricultural domain of CrossVLM-Bench.

Model	Method	Task									
		Det	Count	Ground	Ref-Ground	Seg	InstSeg	RefSeg	ReaSeg	ReCap	Avg
Qwen2.5-VL-7B Yang et al. (2025)	Baseline	2.3	4.3	1.3	13.5	3.7	4.8	20.6	11.1	13.2	8.3
	+SFT	7.6	9.3	13.6	18.4	9.3	11.3	25.9	22.8	22.5	15.6
	+CD-GRPO	**14.6**	**12.0**	**16.3**	**28.5**	**32.1**	**23.4**	**48.1**	**46.3**	**32.8**	**28.2**
Qwen3-VL-8B Bai et al. (2025)	Baseline	5.8	3.8	14.7	22.2	12.8	14.8	39.3	36.9	23.1	19.3
	+SFT	14.6	8.6	32.4	33.2	14.4	17.1	48.1	45.2	32.4	27.3
	+CD-GRPO	**15.6**	**10.7**	**35.8**	**36.4**	**32.1**	**33.1**	**49.7**	**47.0**	**42.8**	**33.7**
MiMo-VL-7B Xiaomi et al. (2025)	Baseline	2.7	5.3	4.5	17.7	8.6	6.5	22.8	22.4	16.1	11.8
	+SFT	4.9	9.3	8.3	23.7	12.2	14.8	29.4	30.2	25.3	17.6
	+CD-GRPO	**13.7**	**12.5**	**11.5**	**26.9**	**29.0**	**26.1**	**34.3**	**40.7**	**35.2**	**25.5**
InternVL3.5-8B Wang et al. (2025)	Baseline	1.4	4.9	10.3	20.1	10.1	8.9	21.1	24.3	11.2	12.5
	+SFT	12.7	11.2	30.2	32.3	18.8	11.1	45.5	45.4	23.8	25.7
	+CD-GRPO	**14.8**	**13.8**	**33.1**	**34.6**	**27.5**	**30.3**	**49.3**	**47.6**	**29.4**	**31.2**

Table 11: Performance on the underwater domain of CrossVLM-Bench.

Model	Method	Task									
		Det	Count	Ground	Ref-Ground	Seg	InstSeg	RefSeg	ReaSeg	ReCap	Avg
Qwen2.5-VL-7B Yang et al. (2025)	Baseline	8.3	16.9	2.4	20.1	14.6	12.3	40.3	32.8	15.8	18.2
	+SFT	27.4	36.6	27.8	33.3	20.1	17.8	50.3	48.6	25.9	32.0
	+CD-GRPO	**34.3**	**42.8**	**34.9**	**36.4**	**26.3**	**24.8**	**64.2**	**63.1**	**33.6**	**40.0**
Qwen3-VL-8B Bai et al. (2025)	Baseline	21.1	14.2	17.8	34.8	18.4	20.8	58.8	46.7	23.4	28.4
	+SFT	54.2	46.4	51.9	58.1	25.2	26.4	84.1	81.6	37.1	51.7
	+CD-GRPO	**55.6**	**49.1**	**52.5**	**60.2**	**69.2**	**64.0**	**85.2**	**82.3**	**42.6**	**62.3**
MiMo-VL-7B Xiaomi et al. (2025)	Baseline	8.6	15.7	5.8	16.5	19.2	15.6	28.1	24.6	17.1	16.8
	+SFT	17.9	31.7	18.3	19.2	21.1	23.4	34.4	26.7	27.2	24.4
	+CD-GRPO	**19.2**	**39.8**	**19.3**	**24.6**	**41.1**	**28.3**	**42.2**	**29.5**	**35.7**	**31.1**
InternVL3.5-8B Wang et al. (2025)	Baseline	14.0	16.2	14.5	25.6	28.8	22.2	43.5	42.1	13.9	24.5
	+SFT	52.6	40.4	52.3	58.3	63.1	28.6	51.2	68.5	23.9	48.8
	+CD-GRPO	**54.5**	**44.1**	**53.9**	**60.6**	**65.2**	**49.2**	**69.8**	**73.3**	**29.4**	**55.6**

C Scaling and Test-Time Adaptation Protocols

The quantitative results of the scaling and test-time adaptation studies are reported in Section 4.5 of the main paper and are not repeated here. Given the main-text space constraints, this section provides detailed protocols for model-capacity scaling, data scaling, in-context learning, and transductive refinement, together with their practical implications.

Common Protocol. Unless otherwise specified, all experiments use the same image-disjoint training and test splits as the main evaluation. Models are adapted independently to each domain using its corresponding training split. Every selected image contributes all nine task annotations, ensuring that changes in data budget do not alter the task composition. SFT is performed for one epoch and followed by one epoch of CD-GRPO using the same optimization and reward configurations described in the main paper. All variants are evaluated on the complete, unchanged test split of the corresponding domain.

Model-Capacity Scaling. We evaluate the 2B, 4B, and 8B variants of Qwen3-VL to isolate the effect of model capacity within a common architecture family. Each model is evaluated before adaptation, after SFT, and after SFT followed by CD-GRPO. The training data, task mixture, prompts, output schemas, domain-specific splits, and evaluation metrics remain identical across model sizes. This controlled design separates the benefit of additional model capacity from that of target-domain adaptation.

Data Scaling. We construct training subsets containing 10%, 30%, and 100% of the images in each domain's training split. Sampling is performed at the image level, and all nine annotations associated with a selected image are retained together. For each data budget, SFT and CD-GRPO use the same selected images and are evaluated on the complete, unchanged test split. This design isolates how the amount of target-domain supervision affects adaptation while preserving the task composition and evaluation protocol.

In-Context Learning. Our in-context learning mechanism constructs a complete pipeline: example selection, dialogue orchestration, and target inference. First, we randomly select five examples from the target domain. The selection criteria require each example to contain an image, a user query, a complete reasoning chain, and a standardised answer; length constraints and structural markers are used to filter out unsuitable candidates. These examples are then formatted as real user–assistant turns, each containing the original image, the text query, and the ground-truth answer. A system instruction and the query for the current target sample are appended at the end, forming a continuous dialogue prompt. The model processes this prompt sequentially, reading the example contexts and transferring their reasoning strategies. Finally, the model produces a prediction for the target sample (bounding boxes, counting results, segmentation descriptions, or language answers) following the same dialogue structure, ensuring consistent output format and significantly improving perception and reasoning accuracy under few-shot conditions. Crucially, the same set of exemplars is reused for inference on every test image within the domain, providing a fixed and consistent context across all queries.

Transductive Refinement. Our transductive learning framework iteratively refines predictions by reusing structured outputs from the previous round across tasks within the same query image. For instance, segmentation inherits previous masks and bounding boxes; detection reuses category boxes and spatial cues; counting and reasoning inject prior numerical estimates; region understanding carries over previous text descriptions. In each new round, we restore all prompts corresponding to a given image to their original coordinates or context, aggregate them into a unified input, and explicitly state in the prompt that the previous outputs may contain quality deviations, asking the model to re-examine and correct them. The model is initialised only once, sequentially reading these task-specific feedbacks to generate new masks, bounding boxes, counting results, or language answers. The latest outputs are then fed back as prompts for the next round, forming a closed loop of "historical output $\rightarrow$ feedback prompt $\rightarrow$ re-reasoning on the target task". This enables the model to accumulate experience and improve performance across multiple tasks through prompt-based self-correction. Importantly, this process never accesses any ground-truth labels from the test split; it relies solely on the model's own predictions and self-generated feedback.

Table 12: **Comparison with data-matched task specialists.** DETR and MaskFormer are adapted on the same target-domain images as CD-GRPO and evaluated on identical test splits using native task metrics.

Task (Metric)	Method	DIOR	NEU-DET	PlantSeg	UODD
Detection (AP)	DETR Carion et al. (2020)	12.7	22.6	6.0	22.4
	CD-GRPO (Ours)	**31.4**	**52.1**	**15.6**	**55.6**
Segmentation (mIoU)	MaskFormer Cheng et al. (2021)	17.6	18.1	11.1	21.8
	CD-GRPO (Ours)	**44.8**	**48.8**	**32.1**	**69.2**

Practical Implications. Taken together, these experiments show that increasing model capacity alone is insufficient to ensure robustness under domain shift; larger models benefit most when coupled with explicit target-domain adaptation. Additional supervision consistently improves performance, while CD-GRPO provides gains across both limited- and larger-data regimes. When parameter updates are impractical, ICL offers lightweight adaptation through a small set of labeled domain examples, whereas transductive learning enables label-free refinement by reusing cross-task predictions from the same test image as structured feedback. These findings suggest a resource-aware strategy that balances model capacity, target-domain supervision, and inference-time computation according to deployment constraints.

D Additional Baseline Comparisons

D.1 Task-Specialist Models

Motivation and Comparison Protocol. CrossVLM-Bench is designed to assess whether a single generative VLM can support heterogeneous perception tasks under domain shift. Nevertheless, it is important to determine whether this unified interface sacrifices performance relative to models specialized for individual tasks. We therefore compare CD-GRPO with DETR Carion et al. (2020) and MaskFormer Cheng et al. (2021), representative specialists for detection and segmentation, respectively. This comparison also controls for access to target-domain data, testing whether the gains of CD-GRPO can be attributed solely to domain-specific supervision.

To ensure a controlled setting, DETR and MaskFormer are initialized from publicly available checkpoints trained on standard natural-image benchmarks such as COCO Lin et al. (2014) and PASCAL VOC Everingham et al. (2010), and are subsequently adapted using exactly the same target-domain images as CD-GRPO. DETR is trained with the source-provided class labels and bounding boxes, while MaskFormer uses the corresponding segmentation masks. Neither specialist has access to additional target-domain images. CD-GRPO uses Qwen3-VL-8B Bai et al. (2025) and the unified nine-task annotations associated with the same training images. All models are evaluated on identical image-disjoint test splits.

Because the specialists do not implement the nine-task generative interface, we compare them only on their native tasks: AP for detection and mIoU for segmentation. For CD-GRPO, detection is evaluated from its decoded boxes, while segmentation follows our standard protocol of converting predicted spatial prompts into pixel-level masks using SAM Kirillov et al. (2023).

Results and Implications. As shown in Table 12, CD-GRPO substantially outperforms the corresponding specialist on both tasks across all four domains. These results show that unified multi-task adaptation need not trade away performance on core geometric tasks: under the same target-image budget, CD-GRPO surpasses dedicated detection and segmentation models on their respective metrics despite supporting nine heterogeneous tasks within a single VLM. The comparison therefore demonstrates that CrossVLM-Bench provides not only broad task coverage but also effective supervision for adapting unified VLMs to specialized perceptual tasks under severe domain shift.

D.2 Domain-Specialist VLMs

Motivation and Comparison Protocol. Domain-specialized VLMs incorporate visual or linguistic priors tailored to a particular application area and may therefore provide stronger initialization

Table 13: **Comparison with domain-specialist VLMs.** Released specialists are evaluated both zero-shot and after post-training on the corresponding CrossVLM-Bench training split following their original adaptation procedures. For RSCoVLM, IAD-R1, and NAUTILUS, CD-GRPO uses the same Qwen2.5-VL-7B backbone; AgriChat retains its original LLaVA-OneVision-7B backbone, while the corresponding CD-GRPO model uses Qwen2.5-VL-7B. All methods are evaluated on identical test splits. Avg. denotes the unweighted mean over nine tasks.

Domain	Model	Det.	Cnt.	Grd.	RefG.	Seg.	ISeg.	RefS.	ReaS.	ReCap	Avg.
Remote sensing	RSCoVLM zero-shot	12.7	13.6	16.3	25.8	23.7	21.7	37.9	39.3	26.8	24.2
	RSCoVLM + post-training	13.8	23.9	17.1	32.7	27.6	26.4	42.8	43.5	28.4	28.5
	CD-GRPO (Ours)	**16.5**	**32.4**	**20.3**	**36.9**	**34.7**	**31.4**	**55.3**	**45.7**	**30.8**	**33.8**
Industrial	IAD-R1 zero-shot	21.9	37.9	21.3	24.7	14.9	17.2	18.1	13.6	20.3	21.2
	IAD-R1 + post-training	22.8	43.1	27.4	22.6	26.9	33.5	39.7	37.2	29.4	31.4
	CD-GRPO (Ours)	**24.6**	**46.8**	**36.9**	**31.0**	**38.8**	**49.1**	**56.6**	**56.5**	**35.9**	**41.8**
Agricultural	AgriChat zero-shot	3.8	5.1	1.9	12.8	4.6	7.9	19.7	14.2	13.8	9.2
	AgriChat + post-training	7.4	9.1	5.5	21.0	12.9	17.5	30.5	28.6	26.7	17.7
	CD-GRPO (Ours)	**14.6**	**12.0**	**16.3**	**28.5**	**32.1**	**23.4**	**48.1**	**46.3**	**32.8**	**28.2**
Underwater	NAUTILUS zero-shot	13.7	23.6	21.9	22.7	17.2	10.1	42.9	29.7	16.8	19.5
	NAUTILUS + post-training	28.5	36.1	30.2	31.5	22.1	19.4	55.8	52.4	28.9	33.9
	CD-GRPO (Ours)	**34.3**	**42.8**	**34.9**	**36.4**	**26.3**	**24.8**	**64.2**	**63.1**	**33.6**	**40.0**

than general-purpose VLMs. We compare against such models to answer two questions: whether their domain-specific knowledge transfers to the unified nine-task setting of CrossVLM-Bench, and whether CD-GRPO remains effective when access to target-domain post-training data is controlled. We select one representative specialist for each domain: RSCoVLM Li et al. (2026a) for remote sensing, IAD-R1 Li et al. (2026b) for industrial inspection, AgriChat Boudiaf et al. (2026) for agriculture, and NAUTILUS Xu et al. for underwater imagery.

We first evaluate each released specialist directly on the corresponding test split. We then post-train it on the CrossVLM-Bench training split following its original adaptation procedure, using our training images and unified nine-task targets. The adapted specialist and CD-GRPO are evaluated on the same image-disjoint test split with identical prompts, output schemas, and task metrics. Thus, both methods have access to the same target-domain images during adaptation.

For comparisons with RSCoVLM, IAD-R1, and NAUTILUS, CD-GRPO uses Qwen2.5-VL-7B Yang et al. (2025) to match the specialist backbone. AgriChat retains its original LLaVA-OneVision-7B backbone; its comparison with Qwen2.5-VL-7B CD-GRPO is therefore matched in model scale and adaptation data, but not in architecture. This distinction is retained when interpreting the agricultural results.

Results. As shown in Table 13, post-training on CrossVLM-Bench substantially improves every domain specialist, confirming that the benchmark provides useful adaptation supervision beyond its diagnostic role. CD-GRPO nevertheless achieves the best result on every task and domain. Compared with the post-trained specialists, it improves the nine-task average by 5.3 points in remote sensing, 10.4 points in industrial inspection, 10.5 points in agriculture, and 6.1 points underwater.

These results separate adaptation effectiveness from access to target-domain data: the specialist models receive the same CrossVLM-Bench training images and are adapted using their original procedures, yet remain below CD-GRPO. Domain-specific pretraining alone is therefore insufficient to support reliable unified perception. Effective adaptation must also align the model with the semantic, geometric, and structural requirements shared by the nine tasks, which is the role served by CD-GRPO.

D.3 Closed-Source VLM Comparison

Motivation and Protocol. Because GPT-4o Islam & Moushi (2025) assists in constructing the language components of CrossVLM-Bench, we further examine whether CD-GRPO merely reproduces the zero-shot behavior of the annotation assistant. Evaluating a closed-source model over all 66,258 image–task pairs would incur substantial API cost. We therefore conduct a controlled case study on a fixed subset randomly sampled from the image-disjoint test split of each domain.

Table 14: **Closed-source VLM case study on randomly sampled test images.** Each selected image is evaluated on all nine tasks. Domain scores are the unweighted nine-task means, and Overall is the mean across four domains.

Method	DIOR	NEU-DET	PlantSeg	UODD	Overall
Qwen3-VL-8B Bai et al. (2025)	14.2	19.5	9.7	24.8	17.1
GPT-4o Islam & Moushi (2025)	21.3	27.6	15.8	33.4	24.5
CD-GRPO (Ours)	**28.7**	**38.2**	**22.1**	**45.6**	**33.7**

Table 15: **General grounding performance before and after CD-GRPO adaptation.**

Method	RefCOCO			RefCOCO+			RefCOCOg	
	testA	testB	val	testA	testB	val	test	val
Qwen3-VL-8B	92.2	85.3	89.9	89.6	77.8	84.5	86.7	86.8
CD-GRPO (Ours)	**93.8**	**88.7**	**92.2**	**91.5**	**82.4**	**87.1**	**88.9**	**89.3**

Every selected image is evaluated on all nine tasks, and the same sampled examples are used for all compared models.

Qwen3-VL-8B Bai et al. (2025), GPT-4o, and CD-GRPO receive identical images, task instructions, and output schemas and are evaluated using the same task-specific metrics. GPT-4o is evaluated directly without target-domain adaptation, whereas CD-GRPO is trained exclusively on the corresponding training splits and has no access to test annotations.

Results. Table 14 shows that GPT-4o improves over the unadapted Qwen3-VL-8B baseline but remains below CD-GRPO in every domain. Although this cost-constrained case study does not constitute a full-benchmark evaluation of GPT-4o, it provides consistent evidence that CD-GRPO is not merely reproducing the annotation assistant's zero-shot behavior. GPT-4o supplies candidate language components during benchmark construction, whereas CD-GRPO learns from trusted source geometry and human-verified annotations through verifiable semantic, geometric, and structural rewards.

E Robustness and Diagnostic Analyses

E.1 General-Capability Retention

Motivation and Protocol. Improving performance in specialized domains should not come at the cost of the broad visual competence acquired during pretraining. To examine this trade-off, we evaluate Qwen3-VL-8B Bai et al. (2025) after CD-GRPO on standard natural-image grounding and multimodal-understanding benchmarks, without additional task-specific tuning. All evaluations follow the original benchmark protocols and compare against the unadapted backbone.

Results. As shown in Tables 15 and 16, CD-GRPO preserves or improves all evaluated general-domain metrics. The gains on RefCOCO Yu et al. (2016), RefCOCO+ Yu et al. (2016), and RefCOCOg Mao et al. (2016) indicate that adaptation does not impair general spatial grounding. Performance also remains stable on MMBench Liu et al. (2024a), OCRBench Liu et al. (2024b) , and RealWorldQA, which cover broader multimodal understanding, text recognition, and real-world visual reasoning. These results show that the improvements in extreme-domain perception are achieved without observable catastrophic forgetting on standard general-domain benchmarks.

E.2 Multi-Run Stability

Motivation and Protocol. To evaluate the run-to-run stability of CD-GRPO, we repeat the complete two-stage adaptation pipeline—one epoch of SFT followed by one epoch of CD-GRPO—with three independent random seeds. The seeds affect data ordering, rollout sampling, and optimization, while the data splits, hyperparameters, prompts, reward formulation, and evaluation protocol remain fixed. The unadapted Qwen3-VL-8B is deterministic under our evaluation setting and is included only as a fixed reference. Run 1 corresponds to the result reported in the main paper, and the aggregate statistics are computed over the three independently trained models.

Table 16: **General multimodal performance before and after CD-GRPO adaptation.**

Method	MMBench	OCRBench	RealWorldQA
Qwen3-VL-8B	78.4	896	71.5
CD-GRPO (Ours)	**78.9**	**898**	**71.8**

Table 17: **Stability over three independent training runs.** The unadapted backbone is included as a fixed reference. Each domain score is the unweighted mean over nine tasks; Overall is the mean across four domains.

Method	Setting	DIOR	NEU-DET	PlantSeg	UODD	Overall
Qwen3-VL-8B	Fixed reference	26.8	28.2	19.3	28.4	25.7
CD-GRPO (Ours)	Run 1 (reported)	41.6	56.2	33.7	62.3	48.5
CD-GRPO (Ours)	Run 2	41.8	56.3	33.6	61.8	48.4
CD-GRPO (Ours)	Run 3	42.3	56.7	34.1	62.2	48.8
CD-GRPO (Ours)	**Mean $\pm$ std**	$\mathbf{41.9 \pm 0.4}$	$\mathbf{56.4 \pm 0.3}$	$\mathbf{33.8 \pm 0.3}$	$\mathbf{62.1 \pm 0.3}$	$\mathbf{48.6 \pm 0.2}$

Results. As shown in Table 17, the three-run mean closely matches the result reported in the main paper, and the standard deviation does not exceed 0.4 points in any domain. This consistently low variance indicates that the gains of CD-GRPO are robust to training stochasticity rather than being driven by a favorable random seed.

E.3 Separating VLM Localization from SAM

Motivation and Protocol. For segmentation tasks, the VLM predicts boxes and points that are subsequently converted into pixel-level masks by the frozen SAM Kirillov et al. (2023). Final segmentation performance therefore reflects both the quality of the VLM-generated spatial prompts and the fixed prompt-to-mask conversion performed by SAM. To identify the source of the improvement from SFT to CD-GRPO, we evaluate localization quality before SAM and mask quality after SAM while keeping the SAM model and inference configuration unchanged.

Table 18 compares point, box, and combined box–point prompting for Qwen3-VL-8B with CD-GRPO. The results are averaged over semantic, instance, referring, and reasoning segmentation within each domain. Table 19 focuses on semantic segmentation and reports three measurements aggregated across domains: box IoU between predicted and reference boxes, normalized spatial similarity between predicted and reference points, and mask mIoU after SAM.

Results. Table 18 shows that boxes and points provide complementary spatial information. Boxes describe object extent, whereas points identify the intended instance or interior region. Their combination achieves the strongest segmentation result in every domain and is therefore adopted as the default prompt configuration.

Table 19 further shows that CD-GRPO improves both pre-SAM localization and post-SAM mask quality relative to SFT. Because the same frozen SAM model and inference procedure are used for both methods, the increase in mask mIoU cannot be explained by changes to the downstream mask generator. Instead, it follows from more accurate boxes and points produced by the VLM. The analysis therefore isolates the source of the improvement between SFT and CD-GRPO to stronger upstream geometric grounding, while holding SAM-based mask generation constant.

E.4 Failure Analysis

Motivation and Protocol. Aggregate task scores do not reveal which failure modes are corrected by CD-GRPO. We therefore conduct a manual error analysis on 100 images sampled from the held-out test splits across the four domains. Every image is evaluated on all nine tasks, yielding 900 predictions per model and 1,800 predictions in total for SFT and CD-GRPO. The same images and queries are used for both models.

We define five recurring error categories: malformed outputs that violate the required response schema; repetition or over-prediction of objects; omission of small or densely distributed objects;

Table 18: **Effect of spatial prompts supplied to SAM.** Results are averaged over semantic, instance, referring, and reasoning segmentation within each domain. Overall denotes the mean across four domains.

Prompt	DIOR	NEU-DET	PlantSeg	UODD	Overall
Point	43.8	55.1	34.9	70.2	51.0
Box	46.1	56.4	38.3	71.5	53.1
Box + point	**48.2**	**60.0**	**40.5**	**75.2**	**56.0**

Table 19: **Pre-SAM localization and post-SAM mask quality.** Box IoU and point similarity evaluate the VLM-generated spatial prompts before SAM, while mask mIoU evaluates the resulting pixel-level masks.

Method	Box IoU	Point similarity	Mask mIoU
SFT	36.3	56.3	27.7
CD-GRPO (Ours)	**38.7**	**88.5**	**48.7**

inaccurate spatial localization; and selection of an incorrect target in referring tasks. The reported rate is the percentage of audited predictions exhibiting the corresponding failure.

Results. As shown in Table 20, CD-GRPO reduces every examined failure mode. The lower rate of malformed responses is consistent with improved format validity, while fewer repeated and excessive predictions indicate stronger control over output structure and cardinality. The reduction in missed small or dense objects further suggests that the model produces more complete object sets. Finally, the substantial decreases in localization and referring-target errors demonstrate improved geometric precision and language–region grounding. Together, these results show that the aggregate gains of CD-GRPO reflect broad improvements in the validity, completeness, and perceptual correctness of the decoded outputs.

F Qualitative Results

We qualitatively compare Qwen3-VL-8B before adaptation, after SFT, and after CD-GRPO on representative examples from the held-out remote-sensing test split. All three models receive the same images and task instructions.

Fig. 8 compares their detection outputs. The unadapted model frequently misses objects, produces poorly localized boxes, or violates the required output structure. SFT improves category recognition and recovers more target objects, but can still produce duplicated, oversized, or spatially implausible predictions. CD-GRPO generates more complete object sets with tighter localization and fewer redundant outputs.

Fig. 9 extends the comparison to counting, referring grounding, and instance segmentation. The unadapted model often returns incorrect counts or associates a referring expression with the wrong instance. SFT reduces these errors but may still omit targets or produce inconsistent spatial predictions. CD-GRPO more accurately preserves object cardinality, follows referring expressions, and localizes the corresponding instances.

For instance segmentation, the visualization displays the boxes and points predicted by the VLM, which serve as intermediate spatial prompts for SAM Kirillov et al. (2023). These prompts are subsequently converted into pixel-level masks for quantitative evaluation; the displayed boxes are therefore not the final segmentation outputs.

These examples complement the quantitative failure analysis in Table 20. While SFT improves target-domain recognition, its predictions may remain incomplete, redundant, or spatially inconsistent. CD-GRPO more reliably aligns the complete decoded output with the object composition and spatial structure of the underlying scene.

Table 20: **Manual analysis of prediction errors.** Each rate is computed over 900 predictions from the same 100 test images for each model.

Error type	SFT (%)	CD-GRPO (%)
Malformed output	8.6	**3.1**
Repetition/over-prediction	12.4	**6.8**
Missed small/dense objects	21.7	**10.2**
Spatial localization error	18.3	**9.5**
Incorrect referring target	15.9	**5.4**

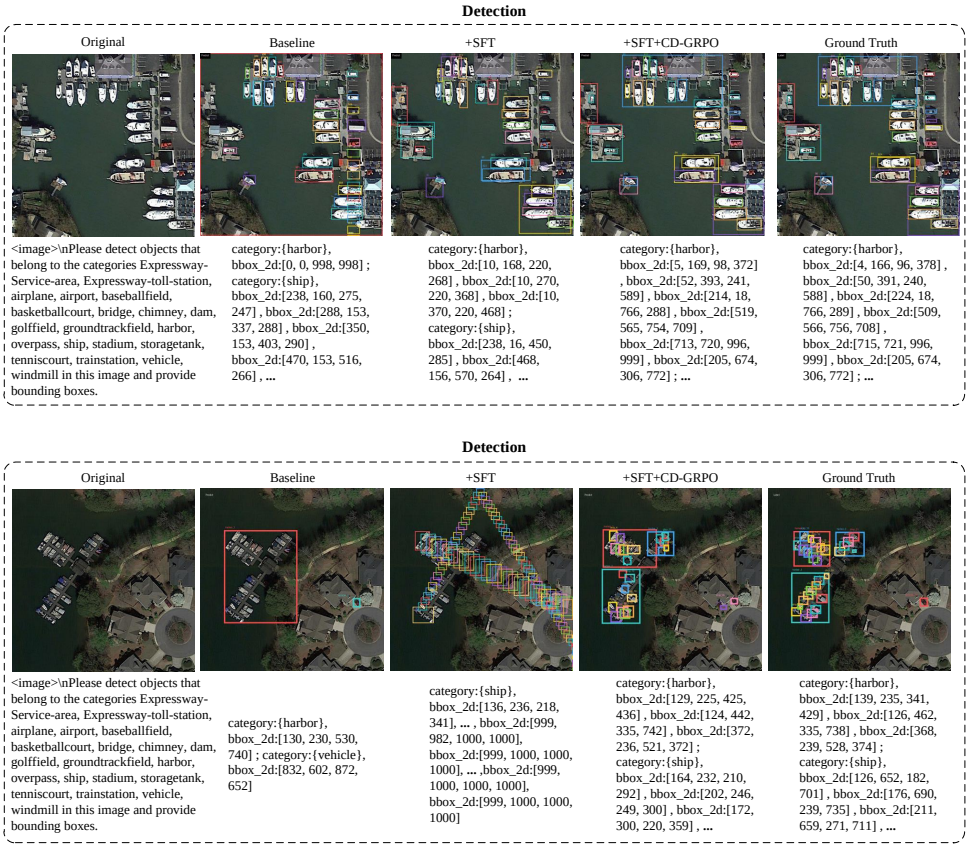

Figure 8: **Qualitative detection comparison on held-out remote-sensing images.** CD-GRPO improves object coverage, localization precision, and output validity relative to both the unadapted model and SFT.

G Limitations and Future Work

CrossVLM-Bench currently covers four image-based extreme domains. Although these domains span substantially different imaging processes, they do not exhaust specialized visual environments or include temporal, multi-image, and expert decision-making tasks. Extending the benchmark to additional sensing modalities and sequential settings is an important direction.

CD-GRPO also relies on task-specific, verifiable accuracy functions. Its reward structure is shared across domains, but introducing fundamentally new or open-ended tasks may require new definitions of correctness. Learning transferable reward models while retaining reliable geometric supervision remains an open problem.

Finally, our segmentation pipeline relies on SAM to convert VLM-predicted boxes and points into pixel-level masks. Although we evaluate both pre-SAM localization quality and the resulting mask quality while holding SAM fixed, the pipeline remains dependent on an external mask generator.



Figure 9: **Qualitative comparison across structured perception tasks.** CD-GRPO improves counting, referring-target selection, and spatial localization. For instance segmentation, the displayed boxes and points are VLM-generated prompts that are subsequently converted by SAM into pixel-level masks.

Developing an end-to-end VLM that directly produces high-resolution masks is beyond the scope of this work and is left for future investigation.