SOLARVQA: A DIAGNOSTIC BENCHMARK FOR VISUAL REASONING IN PHOTOVOLTAIC INSPECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual inspection of a solar cell involves more than detecting a defect. Questions about location, count and covered area require different properties to be combined. We introduce SOLARVQA, a diagnostic visual question answering benchmark with 106,555 questions over 16,322 electroluminescence images. Seven core tasks and six compositional tasks are generated from existing region annotations through image-integrity checks, consistent geometry, explicit ambiguity rules and controls for selected answer shortcuts. Across eleven zero-shot vision-language models, the highest task-macro score remains below a training-fitted nonvisual prior. However, supervised adaptation reveals substantial learnable visual signal. After one epoch of LoRA training, Qwen2.5-VL-7B reaches 64.49% test accuracy, compared with 41.64% before adaptation. With shuffled images, the corresponding scores are 38.02% and 37.08%. The benefit of correct image pairing therefore increases by 21.91 percentage points, with a paired 95% image-cluster interval of [20.86, 23.02]. Compositional performance also improves, although questions that change a simpler answer expose persistent failures. We further show that counting options can inflate ordinal agreement without improving exact accuracy. SOLARVQA supports both learning and diagnosis: its controls identify visual progress while revealing what aggregate gains leave unresolved.

1 INTRODUCTION

Electroluminescence (EL) imaging reveals patterns within photovoltaic cells that support defect inspection. Existing datasets enable models to classify cell images or segment defective regions (Deitsch et al., 2019; Fioresi et al., 2022). A question answering interface asks for more flexible behavior. Given the same cell, a user might ask whether any defect is present, which type occupies the largest area or how many contact defects lie in the upper half. The last question requires several steps: identify the relevant regions, restrict attention to a location and count what remains.

General-purpose vision-language models (VLMs) can accept such questions, but their answers are difficult to interpret without a carefully constructed benchmark. A frequent defect label may be a good guess before the image is inspected. A spatial question may favor the center. A complicated question may retain the same answer as a simpler one. Even the numerical answer choices can reveal enough about scale to make counting errors look small. An evaluation must therefore ask both whether a model selects the correct answer and what evidence contributes to that selection.

We introduce SOLARVQA, an annotation-derived diagnostic benchmark for photovoltaic inspection. It turns existing UCF-EL-Defect (Fioresi et al., 2022) cell images and region labels into 13 multiple-choice tasks. Seven core tasks measure basic properties such as count or covered area. Six compositional tasks combine restrictions and comparisons, allowing us to examine whether a model can answer a more specific question about the same image. Figure 1 shows one complete example with all 13 question types. The benchmark contains 106,555 selected questions across training, validation and test splits.

Reliable questions require consistent treatment of overlapping regions, clipped shapes and nearly tied areas. SOLARVQA combines these construction rules with shortcut controls and explicit evaluation through three contributions:

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.

- **A reproducible inspection benchmark.** We curate source images and annotations, define a common geometry across tasks and retain the information needed to trace each generated target back to its source regions.
- **Compositional tasks with explicit validity rules.** We generate spatial restrictions and cross-property comparisons, reject ambiguous or ineffective cases and balance selected tasks against specific nonvisual shortcuts.
- **An evaluation that separates visual benefit from benchmark priors.** We evaluate eleven zero-shot VLMs and a supervised adaptation baseline, using matched image interventions to measure how learning changes the benefit of visual input. Uncertainty is estimated by resampling images rather than treating their questions as independent.

The central finding is that weak zero-shot performance does not exhaust what can be learned from these images. One epoch of lightweight adaptation substantially improves accuracy and the benefit of receiving the correct image. Yet gains on a compositional task can coexist with failures when its answer differs from the corresponding global question. This combination makes SOLARVQA useful for measuring progress and identifying the capabilities that still need attention.

2 RELATED WORK

Photovoltaic and industrial inspection. ELPV supports cell-level defect assessment (Deitsch et al., 2019), while UCF-EL-Defect provides region annotations for semantic segmentation (Fioresi et al., 2022). We reuse the latter's images and labels to construct question answering tasks. MMAD evaluates multimodal models across industrial anomaly-detection subtasks (Jiang et al., 2025). SOLARVQA focuses on photovoltaic cells and on questions whose targets can be computed from region identity, count, area and position. Its contribution is the task construction and diagnostic evaluation, rather than a new image collection or expert annotation study.

Compositional questions and answer priors. VQA v2 uses complementary images to reduce language priors (Goyal et al., 2017). CLEVR studies compositional reasoning through synthetic scenes with controlled structure (Johnson et al., 2017). GQA uses real-image scene graphs to generate questions and examines their answer distributions and consistency (Hudson & Manning, 2019). SOLARVQA applies this diagnostic approach to specialized inspection imagery. Its targets follow explicit annotation-based operations, while its controls address the priors that remain after question generation. Unlike a held-out-composition benchmark, its splits share the same question grammar.

3 BUILDING SOLARVQA

3.1 CURATING CELL IMAGES AND ANNOTATIONS

The source images are cell crops from module photographs, with VIA-style region annotations (Dutta & Zisserman, 2019). We exclude 480 synthetic images identified by the source dataset's non-`M` filenames. Images without annotation records are excluded rather than assumed defect-free. Jointly empty shape and attribute records identify a clean image within the source annotation scheme.

Two label exclusions define the retained vocabulary. `Unknown` does not specify a defect identity, while `Contact_BeltMarks` occurs too rarely for the intended task suite. The removal audit records 83 images with `Unknown` and 14 with `Contact_BeltMarks`; one contains both. We remove the entire affected image, rather than drop a region and leave a scene that could be mistaken for a complete annotation. The same image-wide policy handles invalid geometry, exact duplicate annotation rows and inconsistent clean/defective records. Malformed region identifiers are repaired by deterministic reindexing.

Deduplication compares decoded RGB pixels and annotation signatures. Equivalent repeats retain one deterministic representative; conflicting annotations for identical images exclude the group. These checks remove exact repeats and contradictory records, but cannot detect all near-duplicates or missing annotations.

The resulting dataset contains 16,322 images, including 11,261 with annotated defects and 5,061 without. A seeded multilabel procedure allocates images to nominal 70/10/20 splits while preserving

support for the nine retained classes. Training, validation and test contain 11,426, 1,632 and 3,264 images respectively. Recorded pixel hashes are unique across splits. Appendix A gives the full removal and split inventories.

3.2 Giving count and area a consistent meaning

A region is a retained source annotation, not necessarily a separate physical defect. Counting measures the number of these regions. Area measures their visible raster masks, clipped to the image boundary. For region set R and mask M_r, the total covered area is

$$A(R) = \left| \bigcup_{r \in R} M_r \right|. \tag{1}$$

Taking the union avoids counting overlapping pixels twice. The same rule applies to a class, a defect family or the whole scene. Different class masks may overlap and are compared separately. Region centers are the mean coordinates of their visible pixel centers, including for clipped shapes.

These definitions matter for spatial questions. A count in the top half assigns a region according to its center. Area in the top half includes only pixels inside that half, even when a region crosses the boundary. More region centers therefore need not mean more covered area. This distinction is the basis of one compositional task.

We exclude close or tied comparisons instead of assigning arbitrary winners. Dominant-class and largest-region comparisons require a relative gap of at least 5% of the larger area. Half-area comparisons require a gap of at least 5% of the combined area. Centroid-based questions exclude centers within 0.5 pixels of a relevant boundary. These are construction rules, not human-calibrated perceptual thresholds. Coverage categories use boundaries of 0.5%, 2.7% and 13.0%, obtained by rounding defective-training coverage quartiles. They describe annotated area, not electrical severity.

3.3 Why compositional questions are included

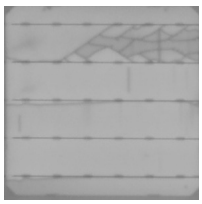
Core questions establish what a model can say about an image as a whole. They ask about defect existence, region count, area-dominant type, total coverage, the location of the largest region, co-occurring classes and the half with greater covered area. Clean images receive only an existence question, so their easy absence answers do not dominate the other tasks.

Compositional questions test whether a model can use those properties under an additional condition or comparison. They serve three purposes. *Restriction* asks about a subset of the scene: the dominant type in a selected half or the number of a specified defect family in that half. *Comparison* combines two properties: whether the half with more family-region centers also has more family area, whether the area-dominant class is also the most frequent class or whether the largest regions of two named classes lie in the same half. *Exclusion* asks how the area comparison changes after removing a named class.

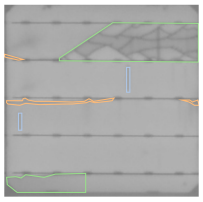
For example, conditional counting first selects Contact regions, then retains centers in the specified half and finally counts them. Excluded-class area first removes all regions of one class, recomputes the remaining union masks and compares their half areas. These questions can reveal failures that a global question misses. However, a restriction need not change the answer. We record whether regional and exclusion targets change relative to their corresponding core target, enabling a separate analysis of the cases where answering globally would be insufficient.

Constructing valid instances. Each compositional question retains its parameters, executable target program and supporting evidence. A seeded axis is assigned before eligibility checks and shared across an image's tasks. Regional halves, conditional families and excluded classes are also assigned before checking validity, without retrying invalid assignments.

Eligibility rules make the operations substantive. Regional questions require defect area both inside and outside the chosen half. Conditional counts require the requested family to exist somewhere and reject filters that leave the entire region set unchanged; a spatially empty result can still be valid. Count/area comparisons require a proper family subset and unique winners. Exclusion must remove visible area while leaving some defect area to compare. Selection keeps at most one question per



Raw image



Annotation overlay

One cell, 13 questions. This selected test image supports every task. Each model call receives the raw image with just one question and its ordered options. Annotation colors indicate defect class: blue for `Contact_FrontGridInterruption`, orange for `Crack_Closed`, and green for `Crack_Resistive`. Bold text below identifies reference answers for the reader.

Task	Question	Ordered answer choices
Core tasks		
Existence	Is there a defect in this image?	A. No; **B. Yes**
Count	How many defect regions are visible in this image?	A. 17; B. 1; C. 4; **D. 7**
Dominant type	Which defect type occupies the largest total visible area?	**A. Crack_Resistive**; B. Interconnect_HighlyResistive; C. Contact_FrontGridInterruption; D. Crack_Isolated; E. Interconnect_Disconnected; F. Contact_NearSolderPad; G. Contact_Corrosion; H. Crack_Closed; I. Interconnect_BrightSpot
Coverage	Which range contains the percentage of the image covered by visible defects?	**A. High: at or above 13.0%**; B. Low: 0.5% to below 2.7%; C. Minimal: below 0.5%; D. Moderate: 2.7% to below 13.0%
Location	In which cell of a 3×3 grid is the center of the largest visible defect region?	A. bottom-right; B. center; C. middle-right; D. middle-left; E. top-center; F. bottom-left; G. bottom-center; **H. top-right**; I. top-left
Co-occurrence	Are two or more distinct defect types visible in this image?	**A. Yes**; B. No
Half area	Which half has the larger total visible defect area?	**A. Top**; B. Bottom
Compositional tasks		
Regional type	Within the top half, which defect type occupies the largest total visible area?	A. Crack_Isolated; B. Interconnect_Disconnected; C. Interconnect_BrightSpot; D. Interconnect_HighlyResistive; E. Contact_NearSolderPad; F. Contact_Corrosion; G. Crack_Closed; **H. Crack_Resistive**; I. Contact_FrontGridInterruption
Conditional count	How many contact defect regions have their centers in the top half?	A. 3; B. 4; **C. 1**; D. 6
Count/area agreement	Comparing the top and bottom halves, does the half containing more crack defect-region centers also have the larger total visible area of those defects?	**A. No**; B. Yes
Dominant/frequent	Does the defect type occupying the largest total visible area also have the most defect regions?	A. Yes; **B. No**
Largest-pair half	Comparing the top and bottom halves, are the centers of the largest visible Crack_Closed region and the largest visible Crack_Resistive region in the same half?	A. Yes; **B. No**
Excluded-class area	Ignoring all Crack_Resistive defects, which half has the larger remaining total visible defect area?	A. Top; **B. Bottom**

Figure 1: A complete example containing all 13 question types. The supplied per-sample visualization is reformatted here using the benchmark wording and ordered options. The annotation overlay and highlighted answers are for visualization only; models never see them. Not every image supports every task. Equivalent per-sample PDFs are included in the anonymous supplementary release as inspection aids.

Table 1: Selected question counts and choice counts K. The last columns show test accuracy (%) of the training-fitted nonvisual prior and the best zero-shot model for each task. The last column is a descriptive envelope, not one model. The division separates core and compositional tasks.

Task	Train	Val.	Test	K	Prior	Best zero-shot
Existence	11,426	1,632	3,264	2	69.0	69.5
Count	7,883	1,126	2,252	4	25.0	30.0
Dominant type	7,775	1,110	2,215	9	41.4	29.7
Coverage	7,883	1,126	2,252	4	25.5	33.6
Location	7,008	999	2,022	9	24.3	24.2
Co-occurrence	7,883	1,126	2,252	2	54.4	66.8
Half area	7,058	1,027	2,030	2	53.0	56.9
Regional type	5,952	863	1,700	9	38.1	28.6
Conditional count	2,848	396	724	4	25.0	29.0
Count/area agreement	672	78	196	2	50.0	54.1
Dominant/frequent	3,371	467	963	2	59.9	59.8
Largest-pair half	3,769	525	1,108	2	51.5	50.5
Excluded-class area	1,117	153	304	2	50.7	54.3

compositional task and six per image, then applies task-specific balancing without refilling discarded slots. Appendix B gives all target operations.

3.4 Controlling specific answer shortcuts

Global-count questions are constructed in blocks of four images with distinct counts. Each image receives the same four numerical choices, independently reordered. All 2,252 test questions belong to complete blocks. Conditional-count questions use complete blocks within each exact question and split; incomplete blocks are discarded. Count/area agreement is balanced between Yes and No within each family and axis. The resulting benchmark contains 81,349 core questions and 25,206 compositional questions, with 74,645, 10,628 and 21,282 in the three splits.

For an exact counting question and an unordered four-option set, every candidate answer occurs equally often. Consequently, any deterministic rule that uses only that question and unordered set achieves exactly 25% accuracy. We verify this for all 271 global-count and 108 conditional-count test groups. The guarantee does not cover option order or other inputs and the numerical choices can still reveal target scale. Other tasks retain answer imbalance. We therefore treat these controls as specific guarantees and measure the remaining priors explicitly.

4 Evaluation protocol

Models and inputs. We evaluate eleven open-weight checkpoints from LLaVA-1.5 (Liu et al., 2024a), LLaVA-NeXT (Liu et al., 2024b), Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025), Idefics3 (Laurençon et al., 2024) and SmolVLM (Marafioti et al., 2025). For the initial comparison, each test question is a separate zero-shot call with an RGB image, the question and lettered choices. The common inspection prompt is passed through the model's native chat template. No annotation, answer or previous question is included. Appendix C reproduces the exact prompt and message structure.

We retain each model's processor, use bfloat16 without quantization and generate greedily with an eight-token cap. Model revisions are pinned. A strict parser accepts a single valid letter, exact choice text or a consistent letter with its choice text. Unparsed responses count as incorrect. All 234,102 original question/model records are complete with no inference errors; only 33 responses are unparsed, all from the two smallest SmolVLM variants.

Scores and nonvisual controls. We report exact accuracy by task and inspect balanced accuracy and class recall where answer imbalance matters. A descriptive suite score gives equal weight to each

task after correcting for uniform guessing:

$$S = \frac{1}{13}\sum_{t=1}^{13} \mathrm{CCA}_t, \qquad \mathrm{CCA}_t = \frac{\widehat{A}_t - 1/K_t}{1 - 1/K_t}. \tag{2}$$

Here K_t is the number of options and $\widehat{A}_t$ is task accuracy. We also report separate core and compositional means. This score corrects uniform chance, not label imbalance. Pooled accuracy weights questions instead of tasks and is reported alongside it. Count MAE and quadratic agreement are secondary diagnostics.

The primary nonvisual prior chooses the most frequent available answer in the training set for the same question and axis. Task frequencies handle unseen groups and frequency ties; remaining ties are resolved lexicographically. The axis is available from the question or half choices. The prior uses no images or question IDs and no test labels are used to fit it. It measures what benchmark-specific answer frequencies can explain. This is deliberately a different information setting from a zero-shot VLM, so losing to the prior does not establish that a model ignores images.

Supervised adaptation. We adapt Qwen2.5-VL-7B, the highest-scoring zero-shot model, using LoRA (Hu et al., 2022) on the language model's attention projections. The vision encoder and all base weights remain frozen. Training uses the 74,645 training questions for one epoch, with rank 16, learning rate 10^{-4} and an effective batch of 32 across two RTX A5500 GPUs. Only the correct option letter and end-of-response token receive loss; images, questions and choices use the original input format. The selected checkpoint maximizes task-macro CCA over a fixed 1,614-question validation subset, sampled without answer labels. Three checkpoints are compared and step 2,333 is selected. This is one training seed with no test-based checkpoint selection. Appendix F gives the complete recipe.

Matched image interventions. Before and after adaptation, we evaluate every test question with its correct image and a shuffled real image while holding wording, ordered options, targets and decoding fixed. Existence questions shuffle across all test images; other tasks shuffle across defective images. Each pool uses one seeded one-to-one donor assignment, shared across model states. Questions from a source image share a donor within a pool. The original zero-shot experiment also includes a gray-image control at source dimensions; no blank-image run is available after adaptation.

The original baseline is reused after matching inference software and obtaining exact raw-response agreement on 104 replay questions per correct/shuffled condition. All four model/image conditions are complete, with no inference errors or unparsed responses. The intervention measures the benefit of correct pairing under this particular replacement scheme. It does not isolate internal reasoning steps (Appendix D).

Uncertainty and consistency checks. We use 5,000 paired image-cluster bootstrap replicates. Each replicate samples images with replacement and retains all questions and conditions belonging to each sampled image. We recompute task accuracies and suite scores, then take percentile 95% intervals. This accounts for shared source images. Ablation intervals are conditional on the saved donor assignment and do not fully capture crossed source/donor dependence or variability across alternative shuffles or training seeds. Task and subgroup intervals are pointwise. Exploratory comparisons of the eleven model scores use whole-image prediction swaps with Holm correction across 55 pairs (Holm, 1979).

Independent rescoring reproduces the original and adapted metrics and the supplied paired intervals. Generation replay recovers all test targets, ordered options and parameters despite small raster differences that leave selection unchanged. Appendix A details these checks.

5 Results

5.1 Zero-shot scores are strongly influenced by answer priors

Among the eleven zero-shot models, Qwen2.5-VL-7B has the highest suite score, $S = 0.079$ [0.061, 0.096], while the nonvisual prior reaches 0.121 [0.104, 0.137] (Table 2). The paired difference is

Table 2: Complete-test results. CCA gives equal weight to tasks after uniform-chance correction. The prior and LoRA use training data; the other model rows are zero-shot. This distinction matters when interpreting the comparison. Intervals resample source images.

Model	Accuracy (%)	Core CCA	Compose CCA	Macro CCA [95% CI]
Training-fitted question prior	43.81	0.146	0.091	0.121 [0.104, 0.137]
Qwen2.5-VL-7B	41.64	0.147	0.000	0.079 [0.061, 0.096]
LLaVA-NeXT-7B	40.07	0.141	-0.010	0.071 [0.054, 0.089]
LLaVA-1.5-7B	39.01	0.108	0.021	0.068 [0.050, 0.085]
Qwen2-VL-7B	38.54	0.098	0.024	0.064 [0.046, 0.081]
LLaVA-1.5-13B	38.41	0.099	0.009	0.058 [0.040, 0.075]
Qwen2.5-VL-3B	38.69	0.123	-0.020	0.057 [0.039, 0.073]
Qwen2-VL-2B	34.78	0.024	0.031	0.027 [0.010, 0.045]
SmolVLM-256M	33.32	0.006	0.011	0.008 [-0.009, 0.026]
Idefics3-8B	32.58	0.001	-0.045	-0.020 [-0.037, -0.003]
SmolVLM-500M	30.25	-0.040	-0.037	-0.039 [-0.056, -0.022]
SmolVLM2-2.2B	30.44	-0.052	-0.029	-0.041 [-0.059, -0.024]
Qwen2.5-VL-7B + LoRA	64.49	0.471	0.305	0.394 [0.377, 0.412]

Table 3: Qwen2.5-VL-7B before and after adaptation, using the same test questions and shuffled-image assignment. Correct images improve accuracy by 4.56 points before training and 26.47 points after training.

Model	Image	Accuracy (%)	Core CCA	Compose CCA	Macro CCA
Zero-shot	Correct	41.64	0.147	0.000	0.079
Zero-shot	Shuffled	37.08	0.057	-0.018	0.022
LoRA	Correct	64.49	0.471	0.305	0.394
LoRA	Shuffled	38.02	0.055	0.039	0.048

-0.042 [-0.067, -0.017]. All zero-shot model point estimates lie below this comparator. The prior also has higher pooled accuracy: 43.81% versus 41.64%. Thus exceeding uniform chance is insufficient evidence of value beyond benchmark priors.

Task-level behavior explains some of this gap. LLaVA-1.5-7B always answers Yes to existence questions, reaching 69.00% accuracy but 50.00% balanced accuracy. The training-mode center gives 24.28% localization accuracy, slightly above every zero-shot model. Qwen2.5-VL-7B selects center on 67.51% of localization questions and has only 13.14% balanced accuracy for dominant type. Nevertheless, its correct-image accuracy exceeds its shuffled-image accuracy by 4.56 percentage points [3.91, 5.21]. A model can benefit from images while still losing to an answer prior.

5.2 Adaptation improves accuracy and the benefit of the correct image

After one epoch, correct-image accuracy rises from 41.64% to 64.49%, a paired gain of 22.85 points [21.97, 23.75]. Task-macro CCA rises from 0.079 to 0.394 [0.377, 0.412]. The adapted model also exceeds the training-fitted prior by 0.274 CCA [0.249, 0.298]. These results establish a stronger supervised baseline without implying a fair training-budget comparison with zero-shot models.

Crucially, most of this improvement depends on correct image pairing (Table 3). Shuffled-image accuracy changes from 37.08% to 38.02%, while the adapted model's correct-versus-shuffled gap is 26.47 points [25.46, 27.52]. To compare the gap directly, define

$$D = (A_{\mathrm{LoRA,correct}} - A_{\mathrm{LoRA,shuffled}}) - (A_{\mathrm{base,correct}} - A_{\mathrm{base,shuffled}}). \tag{3}$$

The paired estimate is $D = 21.91$ points [20.86, 23.02]. For task-macro CCA, the corresponding increase is 0.290 [0.263, 0.318]. All responses parse in both model states. The result supports improved use of the correctly paired visual input rather than an explanation based entirely on answer frequencies or response formatting.

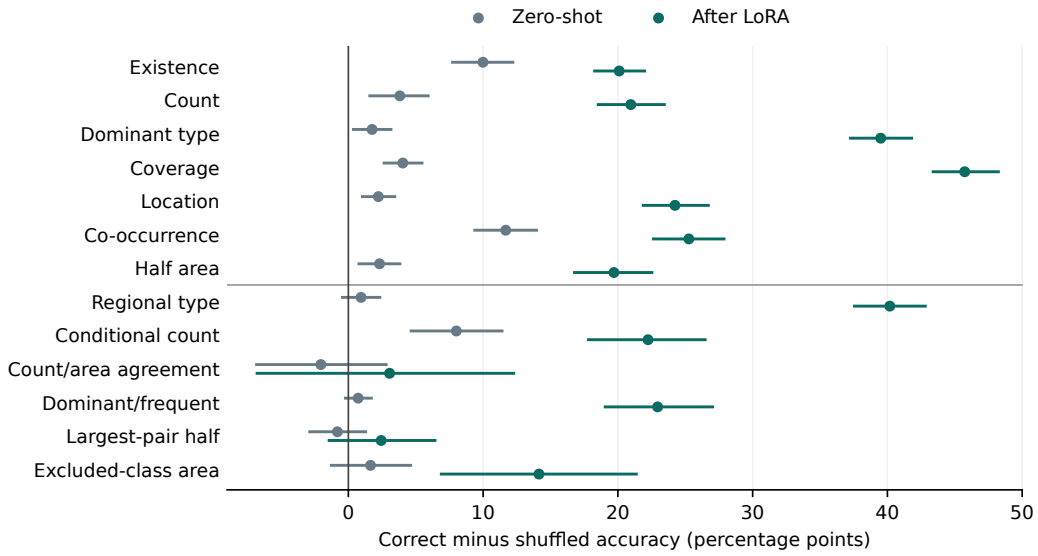


Figure 2: Task-level benefit of correct image pairing before and after LoRA. Points show correct-minus-shuffled accuracy with paired 95% source-image intervals. The division separates core and compositional tasks. Intervals are pointwise and conditional on one training run and one donor assignment.

This improvement extends to the compositional suite. Its correct-image CCA rises from approximately zero to 0.305 [0.271, 0.339]. Its correct-versus-shuffled benefit is 0.267 [0.219, 0.314] after adaptation, compared with 0.018 $[-0.004, 0.040]$ before adaptation. The core-task benefit rises from 0.090 to 0.415. These results provide evidence of learnable image-dependent targets under the current split and question grammar. They do not establish that every individual question is visually unambiguous.

5.3 The benchmark reveals progress and remaining compositional failures

Figure 2 shows broad but uneven visual gains. After adaptation, correct-image accuracy reaches 70.16% for coverage and 69.30% for dominant type, with image benefits of 45.74 and 39.50 points respectively. Dominant-type balanced accuracy rises from 13.14% to 42.55%, so improvement extends beyond the most frequent target. Exact global and conditional counting reach 46.14% and 45.86%, above the 25% reference for rules using only the question and unordered options. Conditional counting retains an image benefit on positive targets: 39.86% correct-image accuracy versus 17.35% shuffled-image accuracy on 582 cases.

Two compositional tasks remain weak. Count/area agreement reaches 52.04% and largest-pair half reaches 53.43%. Their correct-versus-shuffled gains are 3.06 points $[-6.88, 12.38]$ and 2.44 points $[-1.53, 6.53]$. The evidence does not establish an image benefit for either task in this run. Their different supports and target definitions preclude interpreting this as a controlled ranking of reasoning difficulty.

A stronger diagnostic asks whether composition changes the corresponding global answer. Regional-type accuracy reaches 69.65% overall but only 36.76% on the 185 changed-target cases, versus 74.33% on 1,484 unchanged cases. Among the 97 changed cases where the adapted model answers the global question correctly, only 14 also receive a correct regional answer. The original model achieves one such joint success among 41 correct global answers.

Exclusion gives an even sharper warning against reading the aggregate as operation-level competence. Overall accuracy improves to 62.50%, yet it is only 26.67% on the 60 cases where removing a class changes the winning half, compared with 46.67% before adaptation. Accuracy on 222 unchanged cases is 72.52%. These small, target-defined subsets are descriptive diagnostics rather than causal

estimates of a reasoning cost. They show why an image-dependent gain can coexist with failure to carry out the requested restriction or exclusion. Appendix F reports subgroup intervals.

5.4 Counting metrics still require option-only controls

Counting options reveal target scale even when they do not reveal the correct choice. On the global-count test, selecting the third-smallest option gives exactly 25% accuracy but quadratic agreement of 0.755, exceeding every zero-shot model (maximum 0.691). Selecting either middle option gives MAE 3.13, below the best zero-shot point estimate of 3.21.

For balanced blocks with unordered option set S_o, a uniformly chosen option P is independent of target Y given S_o. Nevertheless,

$$\mathrm{Cov}(Y, P) = \mathrm{Var}(\mathbb{E}[Y \mid S_o]), \qquad \kappa_q = \frac{\mathrm{Var}(\mathbb{E}[Y \mid S_o])}{\mathrm{Var}(Y)}. \tag{4}$$

Uniform option sampling has expected agreement 0.623 at 25% accuracy. LoRA improves global-count agreement to 0.885 and MAE to 1.83, alongside its higher exact accuracy. These are useful supervised gains, but the control remains necessary: ordinal agreement alone is not a sufficient measure of solving these multiple-choice counting questions. Appendix E gives the derivation and complete controls.

6 Discussion and limitations

SolarVQA supports a constructive diagnosis of photovoltaic VQA. Its annotation-based programs make targets traceable, its balancing rules address specific shortcuts and its paired image controls test whether training gains depend on the correct visual input. The adaptation experiment demonstrates substantial learnable signal across the suite. Its residual failures also make clear why aggregate accuracy should be accompanied by cases where a compositional operation changes the answer.

Learnability is evidence toward visual answerability, but is not a substitute for independent expert review. The reference targets inherit annotation granularity, specialized class definitions and possible missing regions. No human validation or expert ceiling is available. A model can learn consistent annotation conventions without demonstrating that every target is uniquely recoverable by a human from the raw image.

The experiment uses one adapted model, one training seed and one donor assignment. Intervals condition on the fitted model and assignment. Image replacement can change question applicability. Task populations differ and some subgroups are small. Further seeds and expert assessment would address these limits. Current results do not establish readiness for autonomous quality decisions.

7 Conclusion

SolarVQA provides 13 traceable tasks for learning and diagnosing photovoltaic visual question answering. Zero-shot models struggle against answer priors, while lightweight supervised adaptation substantially improves both accuracy and the benefit of correct image pairing. Target-changed composition cases and option-only counting controls reveal weaknesses that these gains would otherwise obscure. The benchmark therefore offers a measurable path for domain adaptation alongside concrete tests of the capabilities that remain unresolved.

REPRODUCIBILITY STATEMENT

The anonymous supplementary materials include construction audits, analysis code and model revision records. The appendix specifies the target operations, prompt, response parser and statistical procedures. Per-sample visualization PDFs show raw images with annotation overlays and generated questions for inspection; the overlays are never model inputs. Raw intervention responses, donor assignments and the supervised training configuration support reproduction of the reported comparisons. A replay of generation and independent rescoring checks the consistency of targets and results.

ETHICS STATEMENT

This work reuses photovoltaic imagery from UCF-EL-Defect. We do not redistribute the source dataset as a standalone image collection. The supplementary material includes visualization PDFs that embed source images alongside annotation overlays and generated questions for benchmark inspection. Users should obtain the original dataset and associated annotations from the source repository.

AI USE STATEMENT

The authors conducted the model evaluations and adaptation experiments and provided the benchmark implementation and experimental records. OpenAI ChatGPT assisted with literature discovery, experimental and statistical analysis, code preparation, mathematical exposition and manuscript drafting and editing. This included assistance with image-intervention comparisons and the ordinal-counting derivation. Reported statistics were checked against benchmark records and raw predictions through programmatic rescoring. The authors retain responsibility for the interpretation of results and all content of the submission.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, 185:455–468, 2019. doi: 10.1016/j.solener.2019.02.067.

Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the ACM International Conference on Multimedia*, pp. 2276–2279, 2019. doi: 10.1145/3343031.3350535.

Joseph Fioresi, Dylan J. Colvin, Rafaela Frota, Rohit Gupta, Mengjie Li, Hubert P. Seigneur, Shruti Vyas, Sofia Oliveira, Mubarak Shah, and Kristopher O. Davis. Automated defect detection and localization in photovoltaic cells using semantic segmentation of electroluminescence images. *IEEE Journal of Photovoltaics*, 12(1):53–61, 2022. doi: 10.1109/JPHOTOV.2021.3131059.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.00837.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1902.09506.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.09453.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Fei-Fei Li, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017. URL https://arxiv.org/abs/1612.06890.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. URL https://arxiv.org/abs/2408.12637.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. URL https://arxiv.org/abs/2504.05299.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.