A. Limitations and Ethical Considerations

The research reported in this paper aims to refine the capabilities of VLMs by enabling detailed visual understanding without the computational costs associated with uniformly increasing input resolution. This objective aligns with broader goals related to green AI and transparent reporting of the limitations of current VLMs. However, there remain several important considerations for responsible development and deployment.

A first limitation of our study concerns the scope of the experimental setting. Our experiments rely exclusively on English-language models and datasets, which restricts our ability to assess multilingual generalization. Moreover, extending this line of work to multilingual VLMs would require carefully designed mixtures of multilingual VQA data for both training and evaluation.

Another key consideration involves the well-known biases present in VLMs, inherited from large-scale pretraining corpora or from the use of foundation encoders such as CLIP. While we conduct our experiments on publicly available VQA datasets commonly used in prior work, we do not analyze how automatically selected image regions may amplify or suppress particular biases. Caution is therefore warranted before deploying systems such as ours in sensitive real-world environments, and we highlight the need for future research on fairness-oriented evaluation in cropping-based VLM pipelines.

Overall, our approach should be viewed as a step toward more efficient fine-grained visual reasoning, rather than a complete solution to the broader challenges surrounding robustness, fairness, and multilingual accessibility in VLMs.

B. Evaluating the Generated Crops Against Human-annotated Bounding Boxes

The ViCrop authors released human-given bounding box annotations for a subset of the TextVQA dataset, consisting of 4370 samples where each image contains a single annotated region of interest. We made use of this dataset to analyze the bounding boxes produced with **CropVLM**, considering it instead of the VisCoT benchmark since the latter relies on synthetic annotations generated via PaddleOCR for text-centric datasets. Using this data, we calculate the following statistics to describe the behavior of our method, in comparison with ViCrop and UV-CoT. Below, B_p is the predicted bounding box and B_{gt} represents the ground-truth bounding box for each instance.

- **Intersection over Union (IoU):** Measures the relative area of overlap between B_p and B_{gt}.
- **Recall:** Percentage of the B_{gt} area that is covered by B_p.
- **Full Recall:** Percentage of instances where B_p fully contains B_{gt}.
- **Average Relative Bounding Box Size (Size):** Ratio between B_p and the total image area.

Results are presented in Table 1. Notably, bounding boxes generated after GRPO training are significantly larger than those from the SFT stage, regardless of reward type, for **CropVLM** models. This increase in bounding box size, following the reinforcement learning step, suggests that the models learn to capture broader regions to ensure comprehensive information coverage.

A counter-intuitive finding emerges when examining IoU metrics in relation to task performance. IoU does not correlate with TextVQA accuracy when the cropping networks use the same target VLM. For example, among **CropVLM** models at 1024 pixels resolution, the SFT model achieves an IoU of 17.85 but lowest TextVQA performance (53.49), while the LL model has the lowest IoU (13.46) but highest performance (57.14). Similarly, when comparing UV-CoT with ViCrop LLaVA 1.5 methods, higher IoU does not translate to better performance. This suggests that tighter alignment with human annotations does not necessarily correlate with improved question-answering performance. In contrast, recall and full recall metrics, when considered together with relative bounding box size, provide better signals for predicting TextVQA performance within the same answering model.

Additionally, **CropVLM** models trained with reinforcement learning tend to generate progressively smaller bounding boxes as input resolution increases, likely because higher-resolution inputs provide sufficient detail for the model to confidently identify and "zoom in" on specific regions rather than capturing broader areas. This progressive refinement aligns with the expected behavior of an effective cropping mechanism.

C. Benefits at Higher Resolutions

We tested our **CropVLM** model at a 2048×2048 input resolution, pairing it with Qwen 2.5 VL 3B, whose maximum supported resolution is 1792×1792. This setup represents roughly a 16× increase in pixel count compared to the settings used for the results in Tables 4 and 5 of the main paper. The results, shown in Table 2, indicate that **CropVLM continues to provide improvements across most benchmarks**, particularly on datasets that are in-domain with respect to the cropping network's training data. This suggests that visual fragmentation remains an issue even at higher input resolutions, and that our method still helps reduce it. However, the gains are less consistent on V* and HR-Bench 4k/8k, suggesting that for models operating at very high resolutions, the benefits of external cropping may diminish in settings that are out-of-distribution for the cropping model.

Model	Resolution	IoU	Recall	Full Recall	Size	TextVQA
CropVLM SFT	512	14.85	44.16	19.41	14.36	43.26
CropVLM Accuracy	512	15.52	76.34	52.01	26.94	48.41
CropVLM LL	512	10.91	87.95	69.95	45.48	47.55
CropVLM SFT	1024	17.85	51.37	22.56	12.47	53.49
CropVLM Accuracy	1024	18.61	71.95	44.92	16.81	56.05
CropVLM LL	1024	13.46	84.99	64.30	33.39	57.14
CropVLM SFT	2048	17.90	52.41	22.84	13.15	52.16
CropVLM Accuracy	2048	18.07	74.58	47.43	17.83	55.87
CropVLM LL	2048	14.51	84.41	63.56	29.22	56.96
ViCrop LLaVA 1.5 (`rel-attn`)	336	13.70	71.27	46.36	17.20	55.11
ViCrop LLaVA 1.5 (`grad-attn`)	336	13.58	73.45	48.40	18.07	56.07
ViCrop Qwen2.5 VL (`rel-attn`)	448	16.36	55.21	29.08	7.48	72.92
ViCrop Qwen2.5 VL (`grad-attn`)	448	17.74	62.07	33.62	7.44	74.41
UV-CoT	336	14.73	52.61	18.58	10.75	53.33

Table 1. Bounding box quality metrics and corresponding TextVQA performance on a human-annotated subset. CropVLM models are denoted by training stage, where **SFT** represents supervised fine-tuning before reinforcement learning, **Accuracy** represents GRPO performed with accuracy rewards, and **LL** represents GRPO performed with log-likelihood rewards. **Resolution** denotes the input resolution used in each model. ViCrop models are denoted by the model being used and the method. UV-CoT is used here as a cropping network only. TextVQA performance reflects both cropping quality and the underlying answering model (SmolVLM for CropVLM, LLaVA 1.5/Qwen2.5 VL for ViCrop, LLaVA 1.5 for UV-CoT), making direct performance comparisons not meaningful across different answering models.

Method	Reward	TextVQA	DocVQA	InfoVQA	ST-VQA	V*	HR-4k	HR-8k	Average
Qwen 2.5 VL	-	79.14	92.50	75.55	66.67	73.30	**67.75**	63.88	74.11
+ CropVLM	Accuracy	**80.12**	**92.83**	**75.60**	68.71	72.25	65.75	**65.63**	74.41
+ CropVLM	LL	80.07	**92.83**	**75.60**	**68.72**	**74.35**	66.38	63.25	**74.46**

Table 2. Performance of a Qwen 2.5 VL model, operating at 1792x1792 pixels input resolution, paired with **CropVLM**. The **CropVLM** model processes images at 2048x2048 pixels of resolution. The **Average** column shows the average performance across datasets.

D. Ablating the Use of the Full Image

To evaluate the importance of accessing the complete image context when responding to requests, we conducted an ablation study comparing **CropVLM**'s performance when paired with SmolVLM, with and without the full image also being made available, using models trained with accuracy-based rewards and log-likelihood rewards. The results are presented in Table 3. We observe that access to the complete image significantly enhances the model's ability to generate accurate responses, as there is a performance decrease for the version with only crops, across all datasets. Notably, in InfographicVQA, the performance of the model with just a crop still exceeds the baseline model without **CropVLM** for both reward types.

E. Learning to Crop without External Help

Our **CropVLM** models were trained in two stages: initial Supervised Fine-Tuning (SFT) to teach the model to generate valid bounding boxes, followed by GRPO to refine its outputs. To train a model like SmolVLM to generate valid bounding boxes, we first applied SFT using data generated by Qwen 2.5 VL, with slight modifications. Specifically, we expanded smaller bounding boxes to increase the likelihood that they contained the relevant image regions.

To evaluate the impact of external data on **CropVLM**, we repeat the experiments using bounding boxes generated through an alternative method base on exhaustive search. In this setup, we create all possible crops within an $N \times N$ grid for each image, with N equal to 5, and evaluate the log-likelihood of the correct response given the full image, the candidate crops, and the question. SmolVLM, operating at a resolution of 512×512 pixels for efficiency, selects the bounding box corresponding to the highest log-likelihood

Model	Reward	TextVQA	DocVQA	InfoVQA	ST-VQA	Average
SmolVLM	-	55.02	60.13	26.84	58.66	50.16
+ CropVLM (**w/o full image**)	Accuracy	49.13	53.27	27.03	53.93	45.84
+ CropVLM	Accuracy	55.82	61.85	29.76	60.56	52.00
+ CropVLM (**w/o full image**)	LL	53.37	54.72	29.70	56.62	48.60
+ CropVLM	LL	**56.88**	**62.14**	**30.72**	**60.81**	**52.64**

Table 3. Performance of SmolVLM paired with CropVLM, with both models operating at 2048x2048 pixels of input resolution, with and without the full image. The **Average** column reports the average performance across datasets.

as the target for SFT.

Additionally, to ensure the model can produce bounding boxes covering the full coordinate range (0–100), we add random perturbations to the box coordinates. For each training example, noise is sampled uniformly from $[0, \frac{100}{N}]$ and subtracted from the upper-left corner while being added to the lower-right corner, expanding the box outward from its center while keeping all coordinates valid. The reinforcement learning stage is performed as before.

Results are presented in Table 4 demonstrating that **CropVLM** models trained without external bounding box supervision achieve performance similar to those trained with Qwen 2.5 VL supervision. We also observe a larger performance gain from the SFT-only stage to the GRPO stage when supervision does not rely on external data. This underscores both the robustness of our approach and the critical role of reinforcement learning in producing relevant bounding box predictions.

F. Training a Cropping Network with Tight Bounding Boxes

Using a subset of the Visual-CoT training data, which contains tight bounding box annotations for regions of interest in image–question pairs, we train a cropping network to predict regions of interest. This training is performed using supervised fine-tuning with the same hyperparameters as reported in Section 4, without modifying the original bounding box annotations. Specifically, we use the subsets of Visual-CoT corresponding to TextVQA, TextCaps, DocVQA, and InfographicsVQA, totaling 99k samples. In contrast, our GRPO stage used only 62k samples. Table 5 reports VQA performance, while Table 6 reports detection performance on the TextVQA subset described in B.

Results show that a cropping network of this size cannot effectively learn to produce useful bounding boxes for VQA, that tightly enclose the region of interest for each image–question pair. While the model can generate bounding boxes of a similar size, it consistently fails to position them in the correct locations during testing, as reflected in the recall and VQA accuracy metrics. Moreover, even though the **CropVLM - LL** model achieves an IoU similar to that of the Visual-CoT–trained cropping network, its VQA performance is significantly lower: when paired with SmolVLM, it underperforms all our **CropVLM** variants and even the baseline. This further demonstrates that IoU does not correlate strongly with VQA accuracy and should not be used as a proxy metric for VQA performance. Interestingly, our reinforcement-learned **CropVLM** tends to predict substantially larger boxes, suggesting that when precise localization is difficult, the model compensates by expanding the region it selects to achieve higher recall.

G. Expanding the Size of the Human-annotated Bounding Boxes

Datasets with human-annotated bounding boxes, such as the one introduced by the authors of ViCrop, typically provide tight crops around the regions of interest. While precise, these bounding boxes may exclude surrounding visual context that can be crucial for answering certain questions in VQA tasks. To investigate this, we evaluate VQA performance on the TextVQA human-annotated subset under different bounding box expansion factors. As in our main evaluation protocol for other models, we provide both the full image and the corresponding human-annotated crop to the target model. The bounding box centers are kept fixed, while their width and height are adjusted in this test, so that the area becomes a scaled multiple of the original, according to an expansion factor.

As shown in Figure 1, shrinking the bounding boxes immediately reduces performance, confirming that the original annotations are tightly focused on relevant regions. Interestingly, modest expansions can improve performance, suggesting that the tight boxes do not represent an upper bound and that additional surrounding context can be beneficial.

H. CropVLM Qualitative Examples

Figure 2 presents qualitative examples, illustrating the performance of **CropVLM**, when paired with GPT 4.1 nano as the target model, on the out-of-domain V* benchmark. Results are presented for a **CropVLM** model trained using log-likelihood rewards, considering 2048x2048 pixels of input resolution. In turn, Figure 3 presents qualitative

Model	Resolution	External	TextVQA	DocVQA	InfoVQA	ST-VQA	Average
SmolVLM	512	-	39.49	13.68	13.08	47.53	28.45
+ SFT	512	✓	43.55	20.23	14.86	50.39	32.26
+ SFT	512	✗	39.05	21.06	13.65	47.05	30.20
+ SFT + GRPO	512	✓	**47.72**	32.19	**17.32**	**55.29**	**38.13**
+ SFT + GRPO	512	✗	45.83	**33.47**	16.76	53.90	37.49
SmolVLM	1024	-	52.71	47.86	20.12	57.49	44.54
+ SFT	1024	✓	53.46	53.17	21.84	57.71	46.54
+ SFT	1024	✗	52.32	52.20	23.42	57.52	46.36
+ SFT + GRPO	1024	✓	**56.94**	**58.75**	**26.59**	61.26	**50.89**
+ SFT + GRPO	1024	✗	55.97	57.34	26.43	**61.34**	50.27
SmolVLM	2048	-	55.02	60.13	26.84	58.66	50.16
+ SFT	2048	✓	52.29	60.89	26.62	57.37	49.29
+ SFT	2048	✗	53.66	57.71	28.66	58.48	49.63
+ SFT + GRPO	2048	✓	**56.88**	**62.14**	30.72	60.81	**52.64**
+ SFT + GRPO	2048	✗	56.46	59.86	**31.36**	**61.24**	52.23

Table 4. Performance of **CropVLM** paired with SmolVLM across different input resolutions. Both the cropping network and the target model operate at the same resolution. Models marked with ✓ in the **External** column were trained using Qwen 2.5 VL bounding boxes during SFT, while models marked with ✗ were trained using noised crops obtained through exhaustive log-likelihood search. All models undergoing GRPO were trained with log-likelihood rewards. The **Average** column reports the average performance across datasets.

Model	Reward	TextVQA	DocVQA	InfoVQA	ST-VQA	Average
SmolVLM	-	55.02	60.13	26.84	58.66	50.16
+ CropVLM	Accuracy	55.82	61.85	29.76	60.56	52.00
+ CropVLM	LL	**56.88**	**62.14**	**30.72**	**60.81**	**52.64**
+ Visual-CoT	-	50.53	59.25	25.72	56.23	47.93

Table 5. SmolVLM paired with **CropVLM** and the model trained on a subset of the Visual-CoT training data, as denoted by **Visual-CoT**. Here, both the cropping network and the model that responds to the question operate at 2048 pixels resolution. The label **Accuracy** refers to models trained with GRPO where the reward was the accuracy metric for each of the used datasets, while **LL** refers to models trained using the log-likelihood reward of the correct response. The **Average** column shows the average performance across datasets.

examples for a similar setting, but instead considering instances from the TextVQA benchmark. Both cases present success cases, in which GPT 4.1 nano was originally unable to provide the correct answer but the inclusion of an image crop lead to a different result, and also failure cases in which the image crop provided by **CropVLM** distracted the model from providing the correct result.

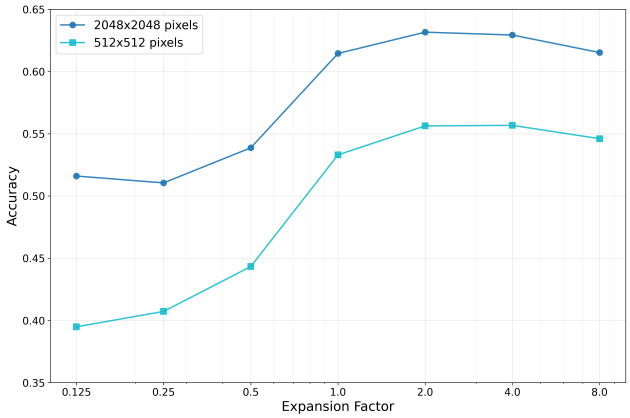

Figure 1. TextVQA performance across multiple bounding box expansion factors, using human-annotated annotations, and with SmolVLM at 512×512 and 2048×2048 input resolutions.

Model	IoU	Recall	Full Recall	Size	TextVQA
CropVLM - SFT	17.90	52.41	22.84	13.15	52.16
CropVLM - Accuracy	18.07	74.58	47.43	17.83	55.87
CropVLM - LL	14.51	84.41	63.56	29.22	56.96
Visual-CoT	14.67	17.23	0.23	1.27	50.24

Table 6. Bounding box quality metrics and corresponding TextVQA performance on a human-annotated subset of the dataset. **CropVLM** cropping networks are denoted by training stage, where **SFT** represents supervised fine-tuning before reinforcement learning, **Accuracy** represents GRPO performed with accuracy rewards, and **LL** represents GRPO performed with log-likelihood rewards. All cropping networks are paired with SmolVLM, and both models operate at 2048x2048 pixels input resolution. **Visual-CoT** denotes a model trained on a subset of the Visual-CoT training data.

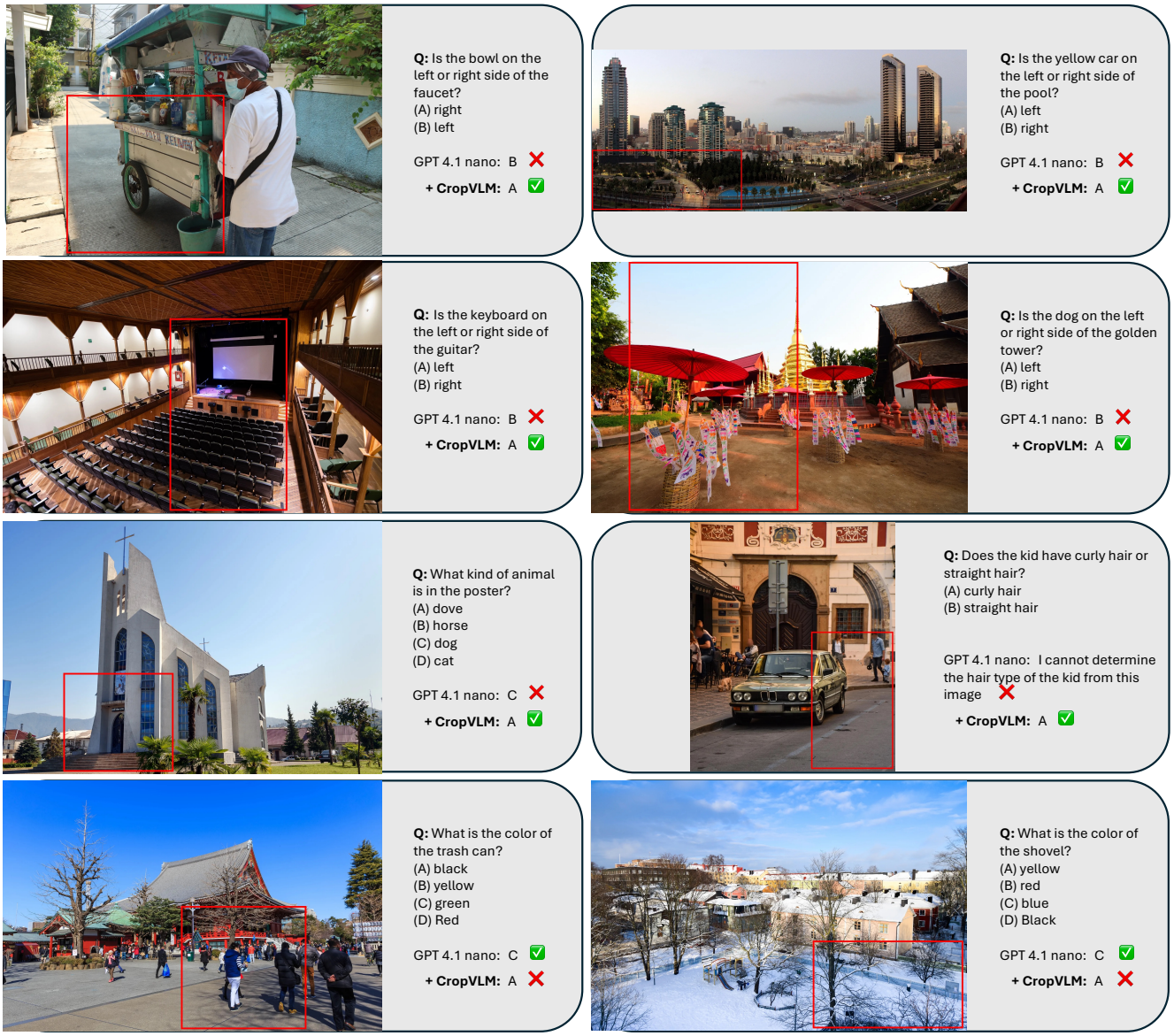

Figure 2. Qualitative examples from the V* Benchmark, where the first 6 cases are successful and the last 2 are failures. Next to each image, we present the question alongside responses from GPT 4.1 nano, and GPT 4.1 nano paired with the **CropVLM** model that accepts images at 2048x2048 pixels of input resolution, and which was trained using log-likelihood rewards. The red bounding box denotes the **CropVLM** proposed region of interest.

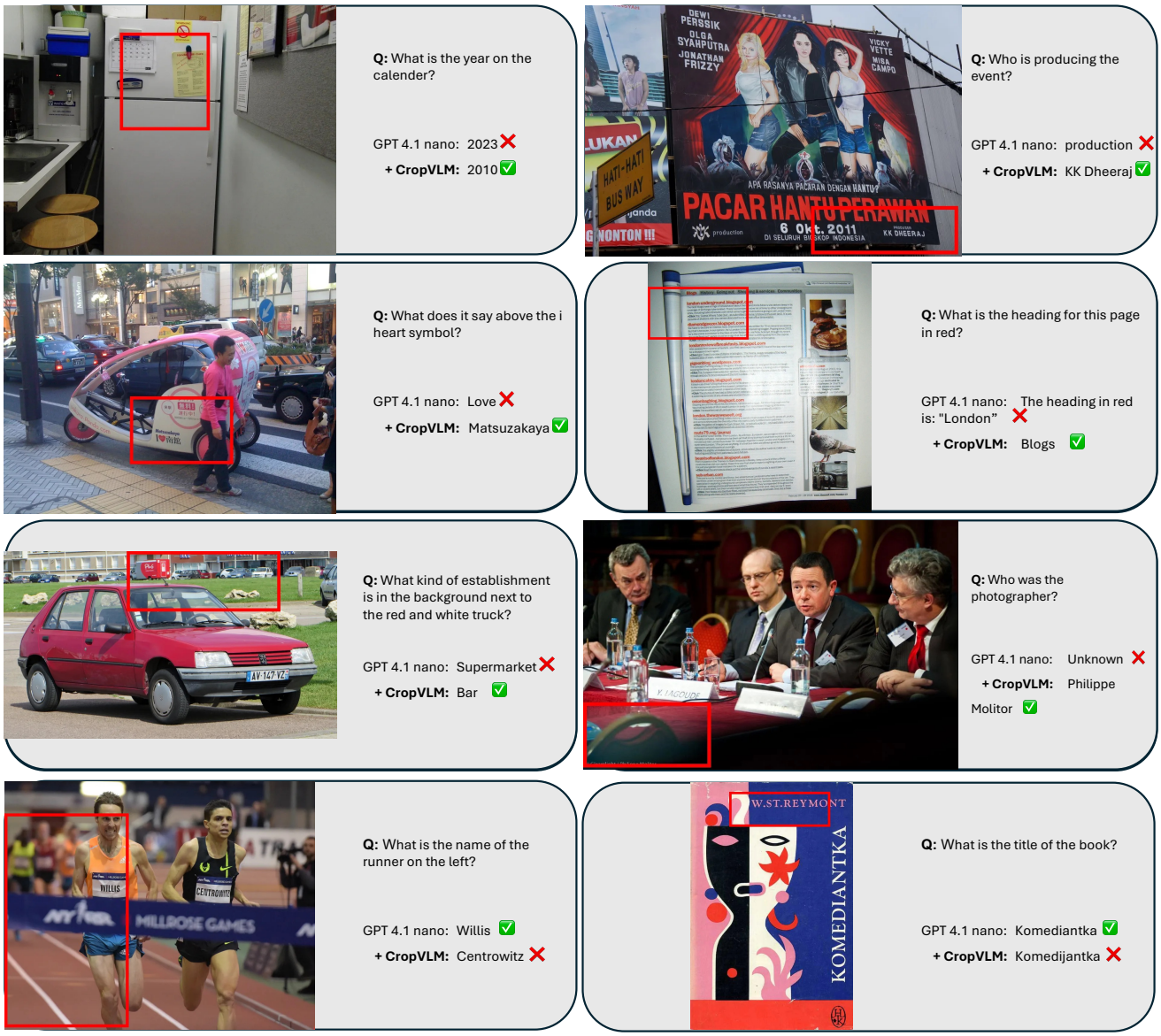

Figure 3. Qualitative examples from TextVQA, where the first 6 cases are successful and the last 2 are failures. Next to each image, we present the question alongside responses from GPT 4.1 nano, and GPT 4.1 nano paired with the **CropVLM** model that accepts images at 2048x2048 pixels of input resolution, and which was trained using log-likelihood rewards. The red bounding box denotes the **CropVLM** proposed region of interest.