Exploring Vision-Language Alignment under Subtle Contradictions

Author
Affiliation
Address
email

Abstract

Vision-language models (VLMs) have made notable progress in tasks such as object detection, scene interpretation, and cross-modal reasoning. However, they continue to face significant challenges when subjected to adversarial attacks. The simplicity of including hidden text in websites points to a critical need for a deeper understanding of how misleading text disrupts performance in multimodal applications. In this study, we systematically introduce faintly embedded and clearly visible contradictory text into a large-scale dataset, examining its effects on object counting, object detection, and scene description under varying text visibility. Our findings show that counting accuracy suffers significantly in the presence of adversarial textual perturbations, while object detection remains robust and scene descriptions exhibit only minor shifts under faint disruptions. These observations highlight the importance of building more resilient multimodal architectures that prioritize reliable visual signals and effectively handle subtle textual contradictions, ultimately enhancing trustworthiness in complex, real-world vision-language scenarios.

1 Introduction

Large Language Models (LLMs) have driven remarkable progress in diverse textual transformation and generation tasks, offering a powerful foundation for emerging multimodal systems (Jiang et al., 2024; Yonekura et al., 2024). Their integration with computer vision architectures has produced vision-language paradigms for applications like image captioning and scene interpretation (Bitton et al., 2023; Liu et al., 2023). Yet, recent work reveals persistent limitations in managing conflicting inputs across modalities, highlighting a need for more robust solutions (Zhao et al., 2023).

Within the realm of vision-language modeling, contradictory textual prompts have become a key concern (Qraitem et al., 2025; Wang et al., 2024; Cheng et al., 2024). An open question focuses on how faintly embedded versus clearly visible contradictory text disrupts the alignment of visual and textual signals. Many vision-language models exhibit performance declines under conflicting cues but lack thorough investigation into subtle contradictions (Qraitem et al., 2025). Addressing these disruptions is essential for applications requiring accurate object recognition, scene understanding, and robust cross-modal integration (Cheng et al., 2024).

This paper systematically explores the influence of both subtle and overt contradictory text by manipulating text visibility in multiple tasks. We address a gap in current benchmarks by isolating the textual component's role in degrading object counting, visual detection, and descriptive accuracy. Novel methodological choices include precise control of text opacity, ensuring that even faint contradictions can alter vision-language representations. These measures illuminate the degrees of visual-linguistic conflict and inform potential avenues for more robust multimodal architectures. Empirical results indicate that contradictory text markedly decreases counting accuracy, dropping by up to 0.078 as text visibility intensifies. Other tasks, such as cat detection, remain comparatively

Preprint.

stable, underscoring the significance of task-specific cues. By comprehensively evaluating how varying text visibility affects system output, this work reveals key vulnerabilities in vision-language alignment. Its contributions include highlighting the need for better handling of misleading lexicon and introducing frameworks that can guide more resilient future VLM designs.

2 Related Works

Vision-Language Models. Vision-language models (VLMs) have attracted considerable attention for their capacity to embed and align textual and visual features, enabling tasks such as image captioning, visual question answering, and object detection (Yonekura et al., 2024; Li et al., 2023). Notable architectures integrate large-scale pre-training to learn joint representations that generalize across multiple modalities (Segal et al., 2022; Yang et al., 2024; Bai et al., 2023; Wang et al., 2023a). Despite rapid advances, these works reveal persistent weaknesses when textual inputs conflict with visual cues, underscoring the need for strategies to handle inconsistent information (Cheng et al., 2024; Qraitem et al., 2025).

Evaluation Metrics and Gaps. Recent efforts propose expanded benchmarks assessing VLMs under varied instructions and adversarial perturbations (Bitton et al., 2023; Wang et al., 2023b; Bai et al., 2023; Dai et al., 2023; Shirnin et al., 2024). However, few approaches systematically manipulate text visibility to uncover the range of model vulnerabilities. Building on these gaps, this paper examines how faint and visible conflicting text affect inference across multiple tasks, contributing a more nuanced evaluation of model robustness in adversarial settings.

3 Methods

We aimed to determine how faintly embedded or clearly visible contradictory textual cues affect a large-scale vision-language model performing visually grounded tasks. Our main hypothesis posited that even subtle contradictions might disrupt object detection, counting, or descriptive accuracy, while more conspicuous text would heighten such disruptions. We were guided by questions around whether the model could discriminate misleading textual information from actual visual cues and how varying degrees of text visibility might alter predictions in tasks such as object enumeration (dogs), object presence (cats), and scene description.

We employed the COCO 2017 training set (Lin et al., 2015), sampling 5000 images to support three tasks: (a) object counting (focusing on dogs), (b) visual search (detecting cat presence), and (c) scene description (identifying objects and colors). Each image was duplicated into three conditions: Original (no text), Faint Text (alpha-blended, near-invisible contradictory text), and Visible Text (clear white font with a black outline). Thus, we aggregated a total of 15,000 image-based data points. We leveraged the `Qwen2.5-VL-7B-Instruct` model (Team, 2024), which processes both images and textual prompts without additional fine-tuning on an NVIDIA A100 GPU. Prompts were customized per task—requesting a count, a yes/no determination, or a compositional scene description.

We gathered performance measures for each task under each condition. For counting, we measured accuracy (perfect dog counts) and mean absolute error (MAE). For visual search, we evaluated accuracy based on correct yes/no recognition of cat presence. The scene description task involved four metrics: object recall, color accuracy, spurious objects, and number of objects mentioned. These metrics offered complementary lenses to understand how contradictory text affects numeric, Boolean, and descriptive outputs.

Alpha-blending in faint text scenarios was carefully tuned so that misinformation was barely visible yet present in the pixel space. In visible text conditions, bold, high-contrast statements were placed in regions of minimal overlap with salient objects. Each modified image was run through the model with standardized prompts, and output parsing was automated to extract dog counts, cat presence, or descriptive text. By systematically varying text visibility while preserving core visual content, we isolated the direct impact of contradictory text on vision-language alignment.

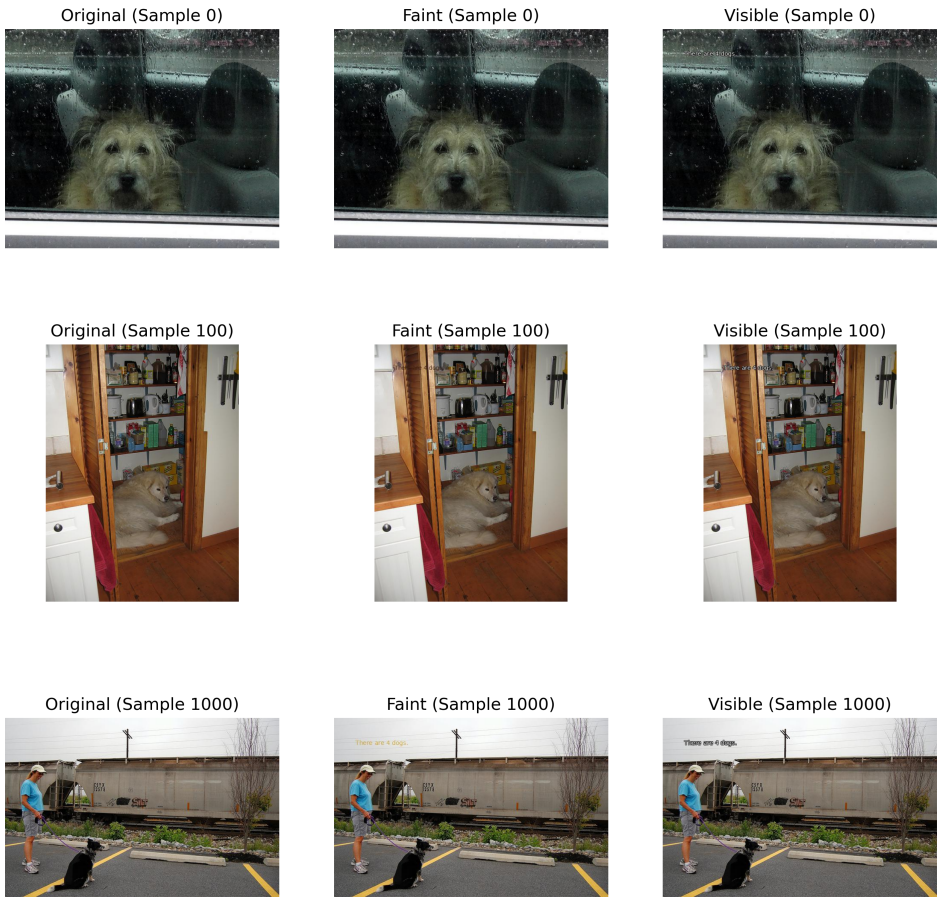

Figure 1: Example input images illustrating the original, faint, and visible versions for different samples used in the study.

4 Results

Contradictory text reduces counting accuracy. Our analysis reveals that the introduction of contradictory text adversely affects the model's ability to count dogs accurately. In the Original condition, the model achieved an exact match accuracy of 0.885. However, when faint contradictory text was added, the accuracy dropped to 0.836, and with more overt (visible) text, it further declined to 0.807. This clear downward trend, also depicted in Figure 2, indicates that textual contradictions can override reliable visual cues. The model appears to become less confident in its numeric predictions when confronted with conflicting information, suggesting that even subtle text-based distractions can significantly undermine counting performance.

Magnitude of counting error increases with text visibility. The disruptive effect of contradictory text is further highlighted by the escalation in Mean Absolute Error (MAE). In the absence of textual interference (Original condition), the MAE was recorded at 0.138. The error nearly doubled to 0.292 under the Faint Text condition and peaked at 0.369 when the text was clearly visible. This marked increase in error magnitude, as shown in Figure 2, underscores how prominently displayed contradictory text not only confounds the model but also leads to increasingly inaccurate numeric

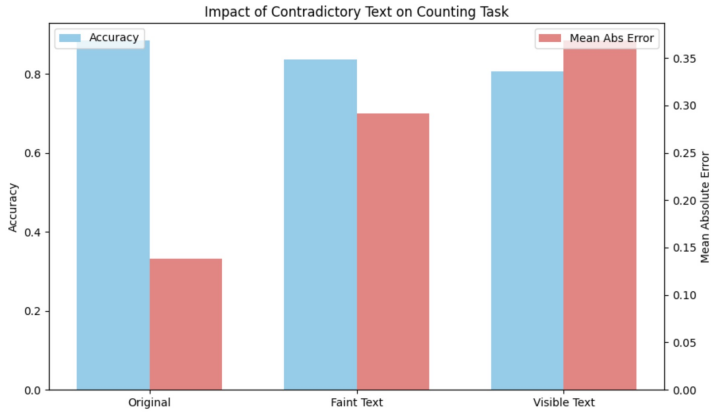

Figure 2: Comparison of dog counting performance for Original, Faint Text, and Visible Text conditions.

predictions. It suggests that as the salience of the conflicting information grows, the model's reliance on precise visual input diminishes.

Object detection remains robust despite contradictions. In stark contrast to counting, the task of detection exhibits remarkable resilience to contradictory text. Across all three conditions—Original, Faint Text, and Visible Text—the accuracy for identifying a cat in an image consistently held at 0.954. Figure 3 illustrates this stability, suggesting that the model relies on highly distinctive visual features that are less susceptible to distraction from textual inputs. This robustness points to the possibility that some visual attributes, such as those critical for cat identification, are deeply embedded in the model's feature extraction process and are therefore minimally impacted by external textual noise.

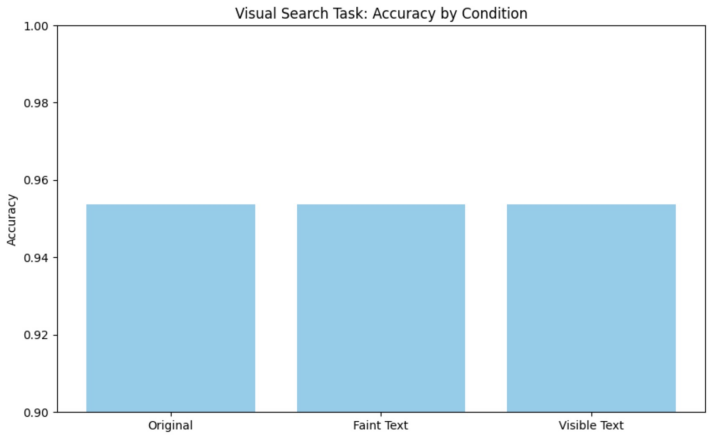

Figure 3: Cat detection accuracies showing no significant changes under faint or visible text.

Faint cues slightly lower object recall in scene descriptions. Beyond object counting, we assessed how contradictory text influenced scene description metrics, with a focus on object recognition. The recall measure, which quantifies the percentage of correctly identified objects, showed a slight decline from 0.555 in the Original condition to 0.539 when faint text was introduced. Although this reduction is minor, it suggests that even subtle textual distractions can hinder the model's ability to fully capture all pertinent objects in a scene. Figure 4 visually illustrates this trend, implying that conflicting information may shift attention away from peripheral visual details.

Color accuracy remains unaffected. Interestingly, the extraction of color attributes appears immune to the influence of contradictory text. The color accuracy metrics remained consistently high, with values of 0.976 (Original), 0.977 (Faint Text), and 0.979 (Visible Text). This near-uniformity implies that color-related features, which are intrinsically tied to the visual composition of an image, are

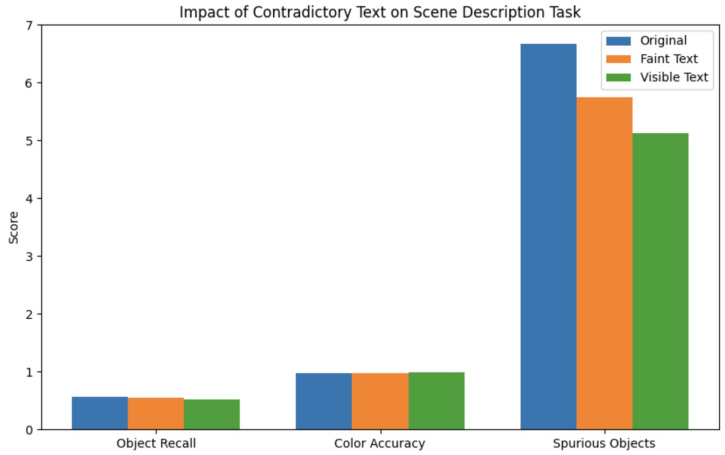

Figure 4: Scene description metrics (object recall and spurious mentions) under Original, Faint, and Visible Text.

robustly encoded by the model. Consequently, even in the presence of distracting textual elements, the model's ability to accurately determine color information remains intact.

Spurious object mentions decrease with contradictory text. An unexpected finding emerged when evaluating spurious object mentions. The model generated an average of 6.67 extraneous object mentions in the Original condition. However, with the addition of contradictory text, these spurious mentions declined to 5.75 in the Faint Text condition and further to 5.13 in the Visible Text condition. This reduction suggests that the model adopts a more conservative approach in its descriptive output when faced with conflicting cues, potentially as a strategy to minimize the propagation of errors. The contradictory text may prompt the model to focus on only the most salient visual elements, thereby reducing the likelihood of over-description.

Overall object mentions also diminish. Complementing the trend observed in spurious mentions, the overall number of objects identified in scene descriptions also decreased under contradictory text conditions. The total count fell from 2.30 in the Original condition to 2.23 with faint text, and further to 2.12 when the text was visible. This contraction in descriptive breadth reinforces the hypothesis that contradictory textual inputs can narrow the model's focus, possibly by diverting attention from less prominent objects. Figure 4 encapsulates these shifts, highlighting how even faint textual distractions can lead to a more limited descriptive output.

5 Conclusion

The findings confirm that textual contradictions can disrupt visually grounded tasks, reinforcing concerns about vision-language alignment (Bitton-Guetta et al., 2023). While counting performance declined, object detection remained stable, suggesting that certain visual features can override misleading text. The models' conservative responses indicate an adaptive recalibration mechanism rather than simple signal merging, which may enhance reliability but hinders precision in tasks like counting. Future work should explore broader contradictory conditions, test diverse models, and refine training strategies that strengthen visual primacy while maintaining flexibility to improve multimodal system resilience in real-world settings.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. 2023.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and L. Schimdt. Visit-bench: A benchmark for vision-language instruction following

inspired by real-world use. 2023.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Y. Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. 2023.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language model, 2024. URL https://arxiv.org/abs/2402.19150.

Wenliang Dai, Junnan Li, Dongxu Li, A. M. H. Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. 2023.

Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2: 1–17, 2024.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A. Plummer. Vision-llms can fool themselves with self-generated typographic attacks, 2025. URL https://arxiv.org/abs/2402.00626.

Elad Segal, Ben Bogin, and Jonathan Berant. Training vision-language models with less bimodal supervision. *ArXiv*, abs/2211.00262, 2022.

Alexander Shirnin, Nikita Andreev, Sofia Potapova, and Ekaterina Artemova. Analyzing the robustness of vision language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2751–2763, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Tiancheng Wang, Yuguang Yang, Linlin Yang, Shaohui Lin, Juan Zhang, Guodong Guo, and Baochang Zhang. CLIP in mirror: Disentangling text from visual images through reflection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FYm8coxdiR.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. 2023a.

Youze Wang, Wenbo Hu, Yinpeng Dong, and Richang Hong. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. *ArXiv*, abs/2308.12636, 2023b.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. Can large multimodal models uncover deep semantics behind images? 2024.

Haruki Yonekura, Hamada Rizk, and Hirozumi Yamaguchi. *Poster: Translating Vision into Words: Advancing Object Recognition with Visual-Language Models*. 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *ArXiv*, abs/2305.16934, 2023.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[D]**

   Explanation: A human selected a "starting paper set" of research papers that the human found interesting. The AI created the hypothesis based on the starting papers. Humans did not provide any feedback.

2. **Experimental design and implementation**:

   Answer: **[D]**

   Explanation: All code and implementation was written by the AI. Humans did not write any code or provide feedback on the implementation.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: No humans provided any feedback or selected any of the data. Humans had no involvement in this portion of the paper.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: Humans modified some citations, a single sentence's phrasing, and placed the AI-generated figures in the final paper. All other parts of the process were completed by the AI.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: It's been really, really great. The limit is only the amount of available compute :)

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Results were verified first with human code-review then by having a human researcher, who is unfamiliar with the AI-generated results, view a summary of the "methodology" section of the paper and re-implement the experiment. The human found the same experimental results as the AI.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: While we do not include a standalone Limitations section, the manuscript acknowledges constraints implicit in our setup. Namely, use of a single VLM, a single dataset and a single synthetic-data generation method, which limits generalization of the results. However, it does not address "computational efficiency of the proposed algorithms and how they scale with dataset size" (quoted from the guidance for limitations) since its not directly relevant for the results of this paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This is an empirical-only paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The methods section specifies the dataset and sample size (COCO-2017; 5k images) and the three experimental conditions (Original / Faint / Visible) and how they are created using the dataset. It identifies the exact model used (Qwen2.5-VL-7B-Instruct) and defines the evaluation metrics for each task (counting: accuracy & MAE; detection: accuracy; etc). These disclosures are sufficient to faithfully reproduce the qualitative main findings that underpin the paper's claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The AI system that wrote the code is proprietary. Right now open-sourcing the code would make it much easier to reverse-engineer the system that wrote the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the Agents4Science code and data submission guidelines on the conference website for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper the details necessary to understand the results. It does clearly state that the experiments use the COCO-2017 training set with a 5,000-image sample, that the Qwen2.5-VL-7B-Instruct model was used without fine-tuning, and detailed descriptions of each tasks and experimental condition. These details are sufficient to understand and discuss the results setup of the experiments. The paper omits low-level information like the sampling strategies used to obtain the 5,000 images, the exact opacity values or font sizes used for the overlay manipulations, and the precise prompt templates, but these are not needed to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports the number of samples in each experimental group and metrics for each task, but does not explicitly compute scores of statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides the GPU compute resources that were used to complete the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: The study is a non-human-subjects, evaluation-only paper that uses an established public dataset (COCO-2017) and does not collect or release new human data, so IRB/fair-wage requirements and privacy/consent risks do not apply. The paper transparently documents its evaluation setup—tasks, conditions, and metrics—supporting scrutiny of any qualitative claims. Taken together, no human subjects, reliance on a well-known dataset, and clear methodological reporting—the paper follows the guidelines as they apply to this type of research.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: While not in a standalone section, the paper does discuss societal stakes on both sides: it highlights risks from adversarially inserted text and the vulnerabilities this creates, and it frames the contribution as improving trustworthiness and guiding more resilient VLMs for real-world deployments."

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.