COREVQA: A Crowd Observation and Reasoning Entailment Visual Question Answering Benchmark

Ishant Chintapatla[†*] **Kazuma Choji**[†*] **Naaisha Agarwal**[†*] **Andrew Lin**[*] **Hannah You**[*] **Charles Duong**[°*]
Kevin Zhu[°*] **Sean O'Brien**[°*] **Vasu Sharma**[°*]

[†]Equal Contribution [*]Algoverse AI Research [°]Equal Senior Authorship

Abstract

Recently, many benchmarks and datasets have been developed to evaluate Vision-Language Models (VLMs) using visual question answering (VQA) pairs, and models have shown significant accuracy improvements. However, these benchmarks rarely test the model's ability to accurately complete visual entailment, for instance, accepting or refuting a hypothesis based on the image. To address this, we propose COREVQA (Crowd Observations and Reasoning Entailment), a benchmark of 5608 image and synthetically generated true/false statement pairs, with images derived from the CrowdHuman dataset (Shao et al., 2018), to provoke visual entailment reasoning on challenging crowded images. Our results show that even the top-performing VLMs achieve accuracy below 80%, with other models performing substantially worse (39.98%-69.95%). This significant performance gap reveals key limitations in VLMs' ability to reason over certain types of image–question pairs in crowded scenes.

[1] [2]

1. Introduction

Vision language models (VLM), such as GPT-4.1 (Achiam et al., 2023) and Gemini 2.5 Pro (Cloud), have shown remarkable capabilities in image understanding and multimodal task completion (Li et al., 2025). As VLMs grow more sophisticated, the demand for rigorous evaluation methods that assess deep visual and textual understanding becomes increasingly critical. (Huang et al., 2023; Agrawal et al., 2018; Goyal et al., 2017).

However, despite offering a wide variety of tasks, existing VLM evaluation benchmarks often fall short in assessing nuanced understanding and deep comprehension of natural situations and analysis, primarily due to simple and easy-to-understand images/questions. These limitations mean models may succeed (and continue to succeed) by exploiting superficial cues or relying on parametric knowledge, without robust visual processing. The impact is impeding VLM improvements due to a scarcity of robust multimodal reasoning assessments (Li et al., 2025; Guan et al., 2024).

To fill this void in VLM assessment, we propose COREVQA (Crowd Observations and Reasoning Entailment Visual Question Answering)–a challenging evaluation benchmark based on images of dense human crowds in complex, natural settings. While existing crowd-based datasets only assess or focus on recognition, detection, and counting (Wang et al., 2020a;b; Sindagi et al., 2020; Wang et al., 2019; Zhang et al., 2016; Xie et al., 2019), COREVQA requires models to integrate fine-grain visual analysis with textual logic in scenarios where visual ambiguity and seemingly trivial, easy-to-miss details are key. COREVQA aims to help researchers spot flaws and gaps in VLM understanding that will spur improvement in robustness.

Our main contributions are as follows.

- We propose a pipeline to synthetically generate difficult questions for specific images based on typical VLM weaknesses.
- We created the first large-scale benchmark with multi-person, crowd-based images for evaluating VLM capability in busy scenarios (using our proposed pipeline).
- We evaluate several state-of-the-art (SOTA) VLMs on COREVQA, revealing across-the-board struggle with nuances and fine details when dealing with images with several people overflowing with diverse shapes, colors, and sizes.

. Correspondence to: Ishant Chintapatla <ishantyunay@gmail.com>, Kazuma Choji <kchoji@hmc.edu>, Naaisha Agarwal <naaishaagarwal@gmail.com>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).

[1]Huggingface link: https://huggingface.co/datasets/COREVQA2025/COREVQA

[2]Github link: https://github.com/corevqa/COREVQA

2. Related Work

2.1. Vision-Language Benchmarks

Several benchmarks have become standard for evaluating core Visual-Question Answering (VQA) abilities. VQAv2 (Goyal et al., 2017) is the successor to the original VQA dataset (Antol et al., 2015), which aims to assess general VQA performance through a more balanced and challenging benchmark than VQA. Though they are still used for standardized evaluation, typical VQA datasets (like OK-VQA (Marino et al., 2019) and TextVQA (Singh et al., 2019)) lack complexity and difficulty. (Goyal et al., 2017). Newer datasets that analyze visual reasoning, understanding, recognition, and question answering include MMTBench (Ying et al., 2024), VCR (Zellers et al., 2019), MM-Vet (Yu et al., 2023), SEEDBench (Li et al., 2023), and NaturalBench (Li et al., 2024). Most recent datasets overlook improving evaluation quality, and focus on assessing a wider range of tasks for easy and standardized comparison with other models (Fu et al., 2024). These benchmarks include MMBench (Liu et al., 2024b), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024), M3GIA (Song et al., 2024). Since most popular VLM benchmarks are designed in a multiple-choice format, further research is required to assess their robustness and reliability (Li et al., 2025). Some others include NTSEBENCH (Pandya et al., 2024) and VLDBENCH (Raza et al., 2025), which target specific areas of visual reasoning. Difficult, targeted datasets like HallusionBench (Guan et al., 2024) have been created to evaluate key VLM weaknesses.

Visual entailment benchmarks such as SNLI-VE (Xie et al., 2019), Defeasible Visual Entailment (Zhang et al., 2025), and VALSE (Parcalabescu et al., 2021) have all created questions that help test the models' ability to understand the text in regards to the image.

Primary crowd-based datasets include NWPU-Crowd (Wang et al., 2020a), JHU-CROWD++ (Sindagi et al., 2020), PANDA (Wang et al., 2020b), and GCC (Wang et al., 2019).

By using a true/false format (rather than the primary usage of multiple choice (Li et al., 2025)) we reduce confounding variables found in answer choices because models cannot pick a correct choice when only given true or false (by nature) (Balepur et al., 2024). Through our crowd-based dataset and synthetic generation we probe both visual entailment and textual comprehension, unlike many datasets mentioned that typically lack complex images (see more in A.6).

3. COREVQA

COREVQA is a novel VQA benchmark designed to evaluate VLMs' capabilities for detailed visual inspection and multi-step visual entailment. The benchmark features true/false statements about images that sound plausible but require careful visual grounding to verify their entailment status.

3.1. Benchmark Overview

COREVQA tests two core capabilities: depth of visual entailment and precision in analyzing fine-grained visual details. Models must classify the truthfulness of complex statements about images, with statements averaging 30 words and featuring complex sentence structures. The binary classification task assesses two main tasks – Meticulous Visual Inspection: identifying subtle details in visual clutter or peripheral regions, and Complex Visual Entailment: understanding spatial relationships, making contextual inferences, and resisting plausible misdirection.

The benchmark has 5,608 unique image-statement pairs. Images come from the CrowdHuman dataset (Shao et al., 2018), featuring diverse indoor and outdoor environments with groups of people. Each image is paired with one unique true/false statement generated through prompting ChatGPT for true statements and Claude for false statements. Ground truths were hand-labeled.

Table 1. Key Statistics of the COREVQA Dataset

Characteristic	Value
Dataset size	5,608 image-statement pairs
true statements	1,566 (27.9%)
false statements	4,042 (72.1%)
Avg. statement length	30.20 words
Statements w/ commas	94.26%

3.2. Data Collection

3.2.1. Image Sourcing

Images were sourced from the CrowdHuman dataset (Shao et al., 2018) (A.7).

3.2.2. True/False Statement Generation

Statements were generated using ChatGPT 4.1 and Claude 3 Opus to increase diversity in the questions (see Appendix A.4 for the reasoning behind this). Both were guided by an iteratively refined 'true prompt' and 'false prompt', designed to create tricky yet answerable statements. The prompt included three key components: (1) a core directive that required generating statements that sounded natural while demanding close visual inspection; (2) content constraints that ensured each statement was grounded in visible elements; (3) a complexity requirement that each statement required chain-of-thought verification. (See Appendix A.1 for the full prompt.)

Statement generation strategies included leveraging subtle

visual cues for true claims and introducing misleading or unsupported elements for false ones. A built-in self-reflection step prompted the model to justify how each statement might deceive, ensuring intentionality and grounded complexity (A.3). Our dataset maintains a slightly unbalanced distribution of true and false statements. (A.2).

3.2.3. Ensuring Question Quality

When labeling the ground truths, we reprocessed any ambiguous questions or made minor grammar edits for clarity.

3.3. Ground Truth

Initially, we attempted to automate ground truth generation using a 'solver prompt' applied to the same models that generated the statements. However, evaluation with a random sample of 250 image-statement pairs revealed only 89% accuracy. To ensure complete accuracy of ground truths, the entire dataset was manually labeled. This ensures that all model scores (accuracy, precision, recall, and f1) are truly representative of their performance on COREVQA.

3.4. Data Analysis

3.4.1. Images

4,927 images (87.9%) from the train01 split and 681 images (12.1%) from the train02 split of the CrowdHuman dataset. These real-world photographs feature groups of people in diverse settings, providing a rich visual foundation for challenging visual entailment statements .

3.4.2. Statements

Statements exhibit significant syntactic complexity, frequent use of contrastive constructions ("while": 32.9%, "despite": 12.7%). Content is people-centric, reflecting the CrowdHuman source, with common terms including "person" (47.3% of statements), "people" (35.4%), and actions like "holding" (46.7%) and "standing" (19.5%).

Over half (57.7%) of statements use spatial terms, 39.0% reference clothing, and 35.1% mention color, highlighting the dataset's focus on detailed visual attributes and spatial understanding.

Notable biases include frequent mentions of specific objects: cameras (17.3%), umbrellas (7.0%), maps (1.1%), and trophies (1.2%). These patterns should be considered when evaluating model performance.

3.5. Dataset Comparison

Table 2 compares COREVQA with other popular VLM benchmarks. While larger datasets like VQAv2 (1.1M samples) provide scale, our benchmark offers a more targeted evaluation through a focused combination of critical features absent in existing benchmarks.

At 5.6K image-statement pairs, COREVQA balances comprehensive coverage with precision-targeted evaluation. This deliberate approach allows for more efficient and meaningful evaluation than would be possible with randomly sampled or user-generated content that test general capability.

COREVQA joins several other datasets in focusing on challenging questions on mulit-person imagery. These include NWPU-Crowd, which only evaluates counting and detection, HallusionBench (Guan et al., 2024), which only focuses on adversarial examples, and SNLI-VE (Xie et al., 2019), which utilizes simple imagery (A.5). COREVQA goes beyond these by providing a dataset with dense visual information and complex visual entailment that requires models to perform multi-step verification.

By strategically combining these dimensions, COREVQA offers unique diagnostic value in assessing VLMs' ability to perform the kind of careful visual verification required in high-stakes real-world applications. Our results reveal that even SOTA models with impressive performance on larger benchmarks struggle significantly when faced with this targeted combination of challenges.

Table 2. COREVQA compared to existing benchmarks

Dataset	Size	Crowd Focus	Adversarial	Fine-grained
COREVQA	5.6K	**Yes**	**Yes**	**Yes**
VQAv2	**1.1M**	No	No	No
SNLI-VE	565K	No	No	Partial
NWPU-Crowd	5K	**Yes**	No	No
HallusionBench	2K	No	**Yes**	No
MMBench	2.9K	No	No	**Yes**
SEEDBench	19K	No	No	**Yes**

4. Results and Analysis

4.1. Experimental Setup

We evaluated GPT-4.1 (OpenAI, 2025), GPT-4o mini (Achiam et al., 2023), Deepseek-Janus-Pro (Chen et al., 2025), LLaVa-NeXT (Liu et al., 2024a), and Qwen2.5 vl 72b (Team, 2025) on COREVQA. All models tested received the same prompt to explicitly respond (to the image and T/F statement) with "True" or "False". We ran all models on the entire dataset, storing their responses. Given the binary classification nature of our task, accuracy is our main evaluation metric. We also report precision, recall, and F1 scores to provide a complete picture of model performance.

4.2. Quantitative Results

GPT4.1 achieves the highest overall accuracy (77.57%) with GPT-4o Mini closely following, demonstrating a reasonable

ability to verify both positive and negative claims. Janus Pro and Qwen2.5 vl 72b also do relatively well (72.31% and 69.95% respectively). However, Janus Pro has significantly low recall and F1 scores, indicating a strong bias toward answering false. LLaVa-NeXT displays near-perfect recall (99.68%) but scores poorly on all other categories (3).

Table 3. Model Performance on COREVQA

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
GPT-4.1	77.57	57.36	76.63	65.60
GPT-4o mini	76.60	56.72	68.45	62.04
Janus Pro	72.31	64.44	1.85	3.60
Qwen2.5 vl 72b	69.95	47.91	87.23	61.85
LLaVa-NeXT	39.98	31.71	99.68	48.12

4.3. Case Studies and Examples

Figure 1 showcases a striking example where all tested models unanimously failed on what appears to be a straightforward visual assessment. Despite the seemingly precise description, all five models incorrectly classified this statement as true when the ground truth is false.

This example illustrates VLMs' tendency to generate plausible but factually incorrect interpretations when precise action recognition is required. The models likely detected people in the foreground and recognized basic poses, but failed to accurately discern if there was exactly one individual who was visibly holding a phone to their ear. See further examples (A.8).

Broader dataset analysis revealed that 21.5% of questions posed significant challenges for models, with at least two models providing incorrect answers. Within these difficult cases, we identified several consistent failure patterns such as not recognizing human actions, overlooking details, inaccurately counting, not understanding spatial relations, or not handling negation properly (A.7.2).

COREVQA demonstrates that fine-grained verification in natural, complex images remains a significant challenge. These patterns reveal fundamental limitations in current VLMs' ability to perform the kind of meticulous visual verification required for real-world applications involving complex scenes.

5. Limitations and Future Work

5.1. Current Limitations

The requirement of human labeling prevents fast scaling. Furthermore, using a true or false format does not provide insight into key areas where a model went wrong (or right) when answering a question. Generating true and false questions solely with ChatGPT and Claude Opus also has the potential to introduce linguistic biases or limit the stylistic diversity of the statements. COREVQA has an uneven split of true and false questions, with there being more false than true.

Figure 1. Case study showcasing a challenging VQA scenario where all models (GPT-4.1, GPT-4o mini, JanusPro, LLaVA-NeXT, and Qwen) responded TRUE when the correct answer was FALSE.

5.2. Suggested Directions for Improvement

Currently we evaluate only one high-parameter open-source model (Qwen) on COREVQA. We can test more models, such as InternVL3-78b, which are high-performing and open-source (Zhu et al., 2025). COREVQA itself could be expanded upon by incorporating crowd data from various sources to increase image diversity. Further analysis and confidence metrics could be conducted to improve the reliability of model accuracy scores. One extension to this paper would be to fine-tune SOTA models on COREVQA and analyze possible improvements in general performance in popular benchmarks. Another extension would be to add non-human-centric images and questions, like groups of animals, to the dataset and analyze model performance.

6. Conclusion

This paper introduces COREVQA (Crowd Observations and Reasoning Entailment), a novel Visual Question Answering (VQA) benchmark designed to rigorously evaluate Vision-Language Models (VLMs). Existing VLM benchmarks often rely on simple images or questions, and existing crowd-based datasets exclusively focus on detection, recognition, and counting. Recognizing this, COREVQA was made with high-quality crowd images and synthetically generated tricky questions, targeting visual entailment capabilities where models must accurately verify or refute claims about image content. Our initial experiments with several state-of-the-art VLMs show that even the most capable models struggle with the complexity of our benchmark. Through COREVQA we aim to expose more gaps in VLM integration of visual understanding and logical analysis, paving the way for further improvements.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Balepur, N., Ravichander, A., and Rudinger, R. Artifacts or abduction: How do llms answer multiple-choice questions without the question? *arXiv preprint arXiv:2402.12483*, 2024.

Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., and Ruan, C. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Cloud, G. Gemini 2.5 pro model. Google Cloud Vertex AI Documentation. URL https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro. Accessed: May 26, 2025.

Fu, C., Zhang, Y.-F., Yin, S., Li, B., Fang, X., Zhao, S., Duan, H., Sun, X., Liu, Z., Wang, L., et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. 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*, pp. 6904–6913, 2017.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Li, B., Lin, Z., Peng, W., Nyandwi, J. d. D., Jiang, D., Ma, Z., Khanuja, S., Krishna, R., Neubig, G., and Ramanan, D. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024.

Li, Z., Wu, X., Du, H., Nghiem, H., and Shi, G. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, 1, 2025.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

OpenAI. Introducing gpt-4.1 in the api, April 2025. URL https://openai.com/index/gpt-4-1/. Announcement of GPT-4.1 model series.

Pandya, P., Gupta, V., Talwarr, A. S., Kataria, T., Roth, D., and Gupta, V. Ntsebench: Cognitive reasoning benchmark for vision language models. *arXiv preprint arXiv:2407.10380*, 2024.

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., and Gatt, A. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.

Raza, S., Vayani, A., Jain, A., Narayanan, A., Khazaie, V. R., Bashir, S. R., Dolatabadi, E., Uddin, G., Emmanouilidis,

C., Qureshi, R., et al. Vldbench: Vision language models disinformation detection benchmark. *arXiv preprint arXiv:2502.11361*, 2025.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

Sindagi, V. A., Yasarla, R., and Patel, V. M. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2594–2609, 2020.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Song, W., Li, Y., Xu, J., Wu, G., Ming, L., Yi, K., Luo, W., Li, H., Du, Y., Guo, F., et al. M3gia: A cognition inspired multilingual and multimodal general intelligence ability benchmark. *arXiv preprint arXiv:2406.05343*, 2024.

Team, Q. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8198–8207, 2019.

Wang, Q., Gao, J., Lin, W., and Li, X. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020a.

Wang, X., Zhang, X., Zhu, Y., Guo, Y., Yuan, X., Xiang, L., Wang, Z., Ding, G., Brady, D., Dai, Q., et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3268–3278, 2020b.

Xie, N., Lai, F., Doran, D., and Kadav, A. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Zhang, C., Kang, K., Li, H., Wang, X., Xie, R., and Yang, X. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia*, 18(6):1048–1061, 2016.

Zhang, Y., Jing, L., and Gogate, V. Defeasible visual entailment: Benchmark, evaluator, and reward-driven optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25976–25984, 2025.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A. Appendix

A.1. Full Generation Prompts

FALSE PROMPT

You are writing a False statement about a given image. The statement should sound completely natural, confident, and plausible — but require close visual inspection to determine if it's actually correct. Your goal is to trick a model that relies on language, assumptions, or common sense — and force it to use detailed visual and contextual reasoning. Do **not** invent things that are fully out of frame, completely occluded, or dependent on unreadable text. Avoid using color as the main trick — models are strong at detecting color. All questions must: - Be complex and require chain-of-thought to answer. The complexity should come from the lengthy and advanced thought process required to answer the image, NOT the detail of the statement itself. **Important Strategy Choices.**: :x: When writing a FALSE Statement: Create a statement that seems plausible but is visually incorrect. The goal is to bait the model into trusting assumptions, not what's actually in the image. Implied Physical Stability — Suggest ambiguous leaning/support just based off proximity, not grounded in visual details. Quantifier Bait — Use vague or tricky counts (e.g. "at least three," "only one") to confuse object detection. Intention/Action Fallacy — Imply

an impossible or contradictory action (e.g. someone trying to do something that seems right, but in context shows is not feasible). Similar Object Confusion — Identify an object that resembles and is used in context of a similar object but is incorrectly labeled. Occlusion Trap — Imply something is visible and confirmable, when it's actually hidden or blocked. Causal Mislead — Imply a cause-effect relationship that isn't actually supported by the image context (e.g. "because X is happening, Y must be true"). Back-to-Front Error — Describe an event in the background as if it's in the foreground, causing false assumptions about prominence or relevance. Non-Functional Object Claim — Attribute an incorrect purpose to an object (e.g. assuming a pipe is a handrail, or a bag is a tool). Time-Based Confusion — Describe a sequence-dependent state that cannot be confirmed from a single frame (e.g. "this person has just finished climbing"). Schema Reversal: Use roles that *sound expected* (e.g. "a coach giving a trophy") but flip them visually (e.g. it's actually a parent or official). Implied Grouping: Suggest affiliation between people (e.g. teammates, family) based on position or dress — even if incorrect. Hidden Contradictions: Embed one small, one or two word, subtle error (e.g. missing ID badge, wrong uniform, incorrect object) inside an otherwise believable sentence. Overloaded Compositions: Combine multiple assumptions (e.g. action + role + object) so the model must validate *each part* to avoid error. — Before responding, analyze your initial response. Make sure it is visually grounded. Also ask, how can I trick someone? How can I trick them into thinking the WRONG answer?

TRUE PROMPT

Generate a true statement about this image that requires sophisticated visual entailment to verify. Format your response as: **STATEMENT:** [Your single true statement about the image] **REASONING:** [Brief explanation of why this statement requires careful visual analysis] The statement should: 1. Be concise and ABSOLUTELY factually verifiable based on visual evidence 2. Employ ONE of these reasoning approaches: - **Spatial reasoning**: Describe precise relationships between multiple elements, their arrangement patterns, or compositional structure - **Temporal/causal inference**: Identify evidence of what just happened or is about to happen based on visual cues of action, intent, or sequence - **Background knowledge integration**: Connect visible details with common knowledge about contexts, functions, or conventions to draw valid conclusions 3. Challenge verification by: - Requiring attention to multiple visual elements simultaneously - Pointing out non-obvious patterns or relationships - Making valid inferences that go beyond listing visible objects 4. Avoid statements that: - Are immediately obvious from a glance - Merely describe the presence of common objects - Could apply to many similar images - Require speculation beyond what's visually evident - Make universal claims about all people or objects in the image IMPORTANT: Double check to make sure each claim in the statement is 100% visually true without exception. When in doubt, use more conservative qualifiers.

A.2. Key Statistics of the COREVQA Dataset

A.3. Prompt Strategies

For true statements, strategies included describing details buried in visual clutter, peripheral elements, subtle interactions, or surprising but correct conclusions. For false statements, strategies included implying unsupported physical states, using confusing quantifiers, mislabeling similar objects, suggesting invisible occluded elements, or creating back-to-front errors. The prompt included a self-reflection step for the LLM to analyze how the statement might deceive, ensuring intentional and grounded trickiness.

A.4. Choosing Models For Question Generation

When initially deciding which model we would use to create questions, we tested Gemini 2.5 Pro, Claude Opus, and GPT-4.1 on a generator prompt for both true and false questions. We ran through random samples of generations from each model and found that true statements from GPT-4.1 and false statements from Opus were most difficult for other models to answer. After this, we built up the generator prompt with patterns and details we saw in these generations.

A.5. COREVQA compared to existing benchmarks

Crowd-Based Visual Complexity COREVQA joins NWPU-Crowd in focusing on challenging multi-person imagery, but with a crucial difference: while NWPU-Crowd addresses only counting and detection, our benchmark evaluates complex visual entailment in these scenes. This requires models to parse dense visual information with multiple subjects, occlusions, and interactions—a fundamental capability for real-world applications.

Adversarial Design Philosophy Like HallusionBench (Guan et al., 2024), our benchmark employs adversarial examples, but we specifically target the intersection of visual complexity and linguistic verification. Each statement is systematically designed to sound plausible while requiring careful inspection to verify or refute. This approach specifically targets the critical gap between language understanding and visual grounding.

Fine-Grained Visual Entailment Most distinctively, COREVQA is the only benchmark specializing in fine-grained visual entailment that demands precise visual

grounding in complex scenes. While SNLI-VE (Xie et al., 2019) offers partial fine-grained verification, it uses primarily simpler imagery and lacks our systematic adversarial design. COREVQA requires models to perform multi-step verification by decomposing complex claims and meticulously verifying each component against visual evidence.

A.6. Further Description of COREVQA

Rather than evaluating various diverse tasks, or exclusively focusing on one aspect of performance, like text recognition or hallucination, COREVQA combines visual entailment and textual comprehension with heavy occlusion from our crowd-based images (Xie et al., 2019). This combination takes the difficult aspects from existing benchmarks and combines them with a focus on crowds to provide quality, in-depth assessment that can be generalizable towards real-world scenarios.

LLMs display an unusual ability to select correct (multiple choice) answers without access to their respective questions, which may reflect the exploitation of dataset artifacts or successful inference from choices alone, rather than a complete understanding of the question and choices together (Balepur et al., 2024). The same capability applies to VLMs because they have an LLM component (Li et al., 2025). By nature, true or false answer choices are insufficient to infer a question from or attach semantic context. Therefore, COREVQA improves upon current benchmarks by removing confounding variables in evaluation.

Though they are high-quality and ideal for evaluation datasets, crowd images have yet to be utilized towards assessing VLM performance outside of recognition, detection, or counting. These datasets either lack questions to go along with their images, or the questions revolve around human detection or counting. Our synthetic data generation pipeline serves as a tool to bring forth new datasets when images are available, but creating questions would be costly or time-consuming. There is currently a lack in high-quality datasets for VLMs for which our pipeline contributes toward a solution.

A.7. Data Analysis

A.7.1. Initial Model Assessment

Our evaluation of five SOTA VLMs on COREVQA reveals significant performance gaps and distinctive response patterns that highlight the challenging nature of our benchmark.

A.7.2. Failure Patterns

Action Recognition Failures: The most prevalent pattern (appearing in 81.3% of difficult cases) involved incorrectly identifying or interpreting human actions and behaviors. Models struggled to distinguish between similar actions or accurately recognize specific gestures, particularly in crowded scenes with multiple subjects. **Detail Oversight:** In 78.1% of challenging questions, models missed crucial visual details essential for accurate assessment. This was especially pronounced in lengthy, detailed descriptions (over 150 characters) that required verifying multiple visual elements simultaneously. **Counting Inaccuracies:** Quantification problems appeared in 60.8% of difficult cases. Models frequently miscounted people, objects, or attributes, particularly when faced with partial occlusions or when distinguishing between similar items was required. **Spatial Reasoning Failures:** Complex spatial relationships (e.g., "left of," "between," "behind") were misinterpreted in 41.7% of challenging cases. Models showed particular weakness in tracking relative positions among multiple subjects, especially when referencing individuals in the background or periphery. **Negation Handling:** Questions involving negation (31.3% of difficult cases) proved consistently problematic. Models struggled to verify statements about what was *not* present or true in an image, suggesting a bias toward confirmation rather than falsification.

Model Performance Performance varied significantly across models on these challenging cases. Janus Pro demonstrated the highest accuracy (64.0%), followed by LLaVA-NeXT (35.6%), while GPT-4.1 and GPT-4o mini performed notably worse (0.3% and 0.2%). This substantial performance gap suggests fundamentally different approaches to visual reasoning among these architectures, with particular implications for handling complex compositional questions about human behavior and interactions.

A.8. Sample of Tricky Questions

The following two questions are some of the most difficult in the entire dataset, where all models got the answer wrong.

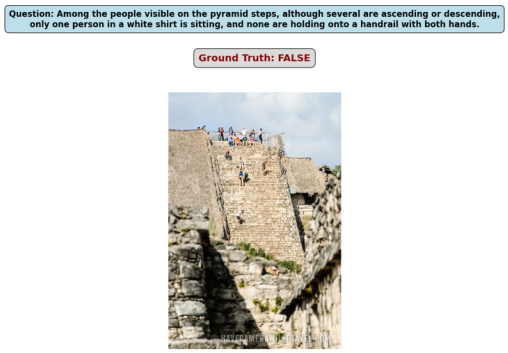

Figure 2. Agreed (Incorrect) Model Answer: TRUE

Figure 3. Agreed (Incorrect) Model Answer: TRUE