Challenges in Visual Entailment for Accessibility

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

In recent years, many benchmarks have been developed to evaluate Vision-Language Models (VLMs) using visual question answering (VQA) pairs, with models demonstrating significant accuracy improvements. However, these benchmarks rarely test visual entailment (determining if an image entails its respective text). Furthermore, existing visual entailment datasets use simple images, which prevent a true evaluation of visual understanding. To address this, we propose COREVQA (Crowd Observations and Reasoning Entailment), a benchmark of 5,608 image and synthetically generated true/false statement pairs. Using images from the CrowdHuman dataset [22], COREVQA provokes visual entailment reasoning in challenging, crowded scenes. Our results show that even 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 the ability of VLMs to semantically understand crowd-based images and reasoning within each image-text pair. This limits VLM capability in assistive technologies for individuals with disabilities. [1] [2]

1. Introduction

Vision-Language Models (VLMs), such as GPT-4.1 [1] and Gemini 2.5 Pro [6], have shown remarkable capabilities in image understanding and multimodal task completion [13]. As VLMs grow more sophisticated, the demand for rigorous evaluation methods that assess deep visual and textual understanding becomes increasingly critical [2, 8, 10]. With VLMs powering assistive technology such as Be My Eyes, thorough evaluation is essential for user safety and quality of life [17].

However, existing VLM evaluation benchmarks often fall short in assessing nuanced comprehension of natural situations, primarily due to their reliance on simple images or questions. These limitations mean that models may succeed by exploiting superficial cues or relying on parametric knowledge without robust visual processing. This scarcity of robust multimodal reasoning assessments impedes VLM improvements [9, 13].

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 focus on recognition, detection, and counting [23, 27–30, 35], COREVQA requires models to integrate fine-grained visual analysis with textual logic in scenarios where visual ambiguity and easy-to-miss details are key. We hope to help researchers spot flaws and gaps in VLM understanding that will spur improvement in assistive robustness and trustworthiness.

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.
- We evaluated several state-of-the-art (SOTA) VLMs on COREVQA, revealing a universal struggle with nuances and fine details when dealing with images overflowing with diverse people, shapes, colors, and sizes.

2. Related Work

2.1. Vision-Language Benchmarks

Several benchmarks have become standard for evaluating core Visual-Question Answering (VQA) abilities. VQAv2 [8], successor to the original VQA dataset [3], aimed to assess general VQA performance through a more balanced and challenging benchmark. Though still used for standardized evaluation, typical VQA datasets (like OK-VQA [16] and TextVQA [24]) often lack sufficient com-

[1]The dataset is available on Hugging Face at: https://huggingface.co/datasets/COREVQA2025/COREVQA.

[2]The demo and generation code are available on Github at: https://github.com/corevqa/COREVQA.

plexity [8]. Newer datasets analyze visual reasoning, understanding, recognition, and question answering, including MMTBench [31], VCR [34], MM-Vet [32], SEEDBench [11], and NaturalBench [12]. Most recent datasets, such as MMBench [15], MMMU [33], MMStar [4], and M3GIA [25], have focused on assessing a wider range of tasks for easier standardized comparison, rather than improving evaluation quality [7]. Other targeted datasets like HallusionBench [9], NTSEBENCH [19], and VLDBENCH [21] have been created to evaluate key VLM weaknesses.

Visual entailment benchmarks such as SNLI-VE [30], Defeasible Visual Entailment [36], and VALSE [20] have all created questions that test a model's ability to understand text in relation to an image. However, these existing visual entailment benchmarks utilize easily understandable images in their assessments, relying on text for entailment. Primary crowd-based datasets include NWPU-Crowd [28], JHU-CROWD++ [23], PANDA [29], and GCC [27].

The original visual entailment task (from SNLI-VE) [30] contains three labels: entailment (if the image contains enough information to conclude the text is true), contradiction, and neutral (if there isn't enough information to conclude). We utilize a true (entailment) or false (contradiction) format, removing the neutral metric from evaluation to provide a more decisive classification task.

Rather than evaluating diverse tasks or focusing exclusively on one performance aspect like text recognition, COREVQA combines visual entailment and textual comprehension with heavy occlusion from our crowd-based images [30]. This combination takes difficult aspects from existing benchmarks and merges them with a focus on crowds to provide a quality, in-depth assessment that is generalizable to real-world scenarios.

3. COREVQA

COREVQA is a novel VQA benchmark designed to evaluate the capabilities of VLMs in 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.

3.1. Benchmark Overview

COREVQA tests two core capabilities: depth of visual entailment and precision in analyzing fine-grained visual details. The binary classification task assesses meticulous visual inspection, which involves identifying subtle details in visual clutter or peripheral regions, and complex visual entailment, which involves understanding spatial relationships, making contextual inferences, and resisting plausible misdirection.

The benchmark contains 5,608 unique image-statement pairs. Images come from the CrowdHuman dataset [22], featuring diverse indoor and outdoor environments with groups of people. Each image is paired with a unique true/false statement generated by prompting ChatGPT for true statements and Claude 3 Opus 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

The images were sourced from the CrowdHuman dataset [22] (see Section 3.3.1).

3.2.2. True/False Statement Generation

After testing several SOTA models, we found that true statements from GPT-4.1 and false statements from Claude 3 Opus were most effective at creating difficult and high-quality questions.

Both models were guided by an iteratively refined prompt designed to create statements that sound natural but require meticulous visual inspection. The prompts included directives for complexity, grounding in visual evidence, and a built-in self-reflection step for the generator to analyze how its statement might trick a model. The exact prompts used for generation are available in our public GitHub repository.

For true statements, the prompt encouraged three main reasoning approaches:

- **Spatial reasoning:** Describing precise relationships between multiple elements. This includes human-to-human interaction, human-to-object interaction, and direction or orientation of moving or still humans and/or objects.
- **Temporal/causal inference:** Identifying evidence of what just happened or is about to happen. Such statements present reasonable inferences based on observations of the situation presented in the given image.
- **Background knowledge integration:** Implementing extensive details about the background of a scene in the statement challenges models to verify all parts of the image.

For false statements, the prompt employed a range of adversarial strategies designed to exploit common VLM weaknesses. These included:

- **Occlusion Trap:** Implying something is fully visible when it is actually partially or fully hidden.
- **Causal Mislead:** Suggesting a cause-and-effect relationship not supported by the visual context (e.g., "because X

is happening, Y must be true”).

- **Schema Reversal:** Flipping expected social roles (e.g., describing a parent handing a trophy as a coach).
- **Quantifier Bait:** Using counts (e.g., “at least three,” “only one”) for simple object detection. Generally, these statements mention detailed attributes of the objects to throw off VLMs and make them doubt their count.
- **Hidden Contradictions:** Embedding a single, subtle error (e.g., a missing ID badge or an incorrect object) within an otherwise believable sentence.

This systematic approach ensures that the benchmark’s difficulty stems from intentional, grounded complexity rather than random chance.

3.2.3. Quality Control and Ground Truths

All ground truths were manually labeled to ensure complete accuracy. While labeling, we also reprocessed any ambiguous statements or made minor grammatical edits for clarity.

3.3. Data Analysis

3.3.1. Images

The dataset includes 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.

3.3.2. Statements

The statements exhibit significant syntactic complexity, with frequent use of contrastive constructions (”while”: 32.9%, ”despite”: 12.7%). The 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%).

More than half (57.7%) of the 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.

3.4. Dataset Comparison

Table 2 compares COREVQA with other popular VLM benchmarks. Our dataset joins several other datasets in focusing on challenging multi-person imagery. These include NWPU-Crowd, which only evaluates counting and detection, HallusionBench [9], which only focuses on adversarial examples, and SNLI-VE [30], which uses primarily simpler imagery. 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 a unique diagnostic value in assessing the ability of VLMs to perform the kind of careful visual verification and reasoning required in real-world applications.

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 [18], GPT-4o mini [1], Deepseek-Janus-Pro [5], LLaVa-NeXT [14], and Qwen2.5 vl 72b [26] on all statements of COREVQA. All models were given the same prompt to explicitly respond with ”True” or ”False”. Our primary evaluation metrics are accuracy, precision, recall, and F1. We also introduce failure patterns to assess areas of challenge within each statement.

4.2. Quantitative Results

GPT-4.1 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 perform 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 metrics (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. Failure Patterns

21.5% of questions were particularly challenging, with at least two models providing incorrect answers.

Categorization using VLM(s) as a judge: We used the same ChatGPT and Claude models used for generation as judges to categorize these difficult questions into the five categories mentioned below. If a statement was answered incorrectly, it is considered a failure of all respective categories attributed to that statement (there can be multiple). Both models were given similar prompts as were used for generation, and were given further instructions for categorization.

Anytime the models disagreed on a categorization, humans were used to select the best-fitting categories. Among the 1208 questions, conflicts were present in only 48 (3.97% disagreement). This includes cases where one of the models attributes more categories to the given statement than the other.

To test the reliability of this VLM-as-a-judge approach, we selected a random sample of 50 statements (127 VLM categorizations), and found an accuracy of 96.06% where human categorizations of those same questions were ground truths (122/127).

Action Recognition Failures: (81.3% of difficult cases) Models often failed to understand complex human actions, or contextual behaviour (e.g., "a person is actively hailing a cab").

Detail Oversight: (78.1%) This pattern highlights a core challenge in visual grounding. Models struggled to verify multiple, disparate visual facts asserted in a single, long statement.

Counting Inaccuracies: (60.8%) These are indicated by failures in quantification, especially in occluded scenes. Model predictions below and above the ground truth were both prominent.

Spatial Reasoning Failures: (41.7%) Models frequently misinterpreted complex spatial prepositions like "between," "behind," or "to the left of," particularly when the statement involved multiple subjects.

Negation Handling: (31.3%) By nature of the images and statements, it is often more demanding to verify something's presence than to confirm its absence. This includes statements such as "no one is wearing a hat".

4.4. Case Studies and Examples

Figure 1 showcases an example where all tested models unanimously failed. The statement requires careful application of several reasoning steps: counting ("only one"), action recognition ("holding a phone to their ear"), and negation ("no one...is both carrying an umbrella and wearing a hat"). This statement is a case of detail identification, negation handling, and action recognition.

5. Limitations and Future Work

5.1. Current Limitations

The requirement of human labeling prevents fast scaling. Furthermore, generating questions solely with ChatGPT and Claude Opus has the potential to introduce linguistic biases or limit the stylistic diversity of the statements. In a binary format, VLMs can attain non-trivial (50%) accuracy through random guessing. Another limitation is that when a model responds falsely, we cannot confirm which part of the

Figure 1. Statement: Among all people crossing the street, only one is visibly holding a phone to their ear while walking, while no one in the scene is both carrying an umbrella and wearing a hat. Ground truth: FALSE. All models (GPT-4.1, GPT-4o mini, JanusPro, LLaVA-NeXT, and Qwen) responded TRUE.

statement the VLM believes is false. Finally, COREVQA contains an uneven split of true and false statements.

5.2. Suggested Directions for Improvement

Future work should test more models, such as InternVL3-78b [37], which are high-performing and open-source. Incorporating crowd data from various sources (like non-human images) would increase the generalizability of COREVQA. Further analysis and confidence metrics could be conducted to improve the reliability of model accuracy scores. Finally, COREVQA could be used for finetuning VLMs, to evaluate potential performance improvements in general visual and textual tasks.

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, while existing crowd-based datasets exclusively focus on detection, recognition, and counting. Recognizing this gap, COREVQA was created with high-quality crowd-sourced images and synthetically generated challenging statements, targeting visual entailment capabilities where models must accurately verify or refute claims about image content. Our experiments identified under 80% accuracy from state-of-the-art VLMs. Through COREVQA, we aim to expose gaps in assistive VLM technology in real-world scenarios.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 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*, pages 4971–4980, 2018. 1

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 2

[5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 3

[6] Google Cloud. Gemini 2.5 pro model. Google Cloud Vertex AI Documentation. Accessed: May 26, 2025. 1

[7] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*, 2024. 2

[8] 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*, pages 6904–6913, 2017. 1, 2

[9] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, 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*, pages 14375–14385, 2024. 1, 2, 3

[10] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 1

[11] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2

[12] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024. 2

[13] Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, 1, 2025. 1

[14] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3

[15] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1

[17] OpenAI. Be my eyes, 2024. Accessed: August 23, 2025. 1

[18] OpenAI. Introducing gpt-4.1 in the api, 2025. Announcement of GPT-4.1 model series. 3

[19] Pranshu Pandya, Vatsal Gupta, Agney S Talwarr, Tushar Kataria, Dan Roth, and Vivek Gupta. Ntsebench: Cognitive reasoning benchmark for vision language models. *arXiv preprint arXiv:2407.10380*, 2024. 2

[20] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021. 2

[21] Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, et al. Vldbench: Vision language models disinformation detection benchmark. *arXiv preprint arXiv:2502.11361*, 2025. 2

[22] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1, 2

[23] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2594–2609, 2020. 1, 2

[24] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1

[25] Wei Song, Yadong Li, Jianhua Xu, Guowei Wu, Lingfeng Ming, Kexin Yi, Weihua Luo, Houyi Li, Yi Du, Fangda Guo, et al. M3gia: A cognition inspired multilingual and multimodal general intelligence ability benchmark. *arXiv preprint arXiv:2406.05343*, 2024. 2

[26] Qwen Team. Qwen2.5-vl, 2025. 3

[27] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207, 2019. 1, 2

[28] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020. 2

[29] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278, 2020. 2

[30] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 1, 2, 3

[31] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 2

[32] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[33] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, 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*, pages 9556–9567, 2024. 2

[34] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 2

[35] Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, Rong Xie, and Xiaokang Yang. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia*, 18(6):1048–1061, 2016. 1

[36] Yue Zhang, Liqiang Jing, and Vibhav Gogate. Defeasible visual entailment: Benchmark, evaluator, and reward-driven optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25976–25984, 2025. 2

[37] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4