

# SafeHatQA: Vision-Language Model for Hard Hat Detection

**Eleanna Panagiotou**

University of Wisconsin-Madison  
Madison, WI, USA  
panagiotou@wisc.edu

**Zack Sifakis**

University of Wisconsin-Madison  
Madison, WI, USA  
zsifakis@wisc.edu

## 1 Introduction

The construction industry is one of the most hazardous sectors, with workers frequently exposed to serious risks, including head injuries from falling objects. These injuries can lead to severe consequences, including long-term disability and fatalities. Such injuries also result in significant economic costs and highlight the critical role of safety compliance in protecting worker welfare. In 2012 alone, over 65,000 construction-related head injuries required days off work, and head injuries accounted for 1,020 fatalities, as reported in the National Safety Council’s Injury Facts chartbook (2015) (HexArmor, 2019). Additionally, a study by the Korea Occupational Safety and Health Agency in 2016 found that 41.2% of reported injuries in the construction industry affected the head (Kim et al., 2018).

This alarming data emphasizes the need for effective safety compliance mechanisms, with helmet detection playing a critical role in reducing injury risks (Li and Li, 2020; Mahalakshmi et al., 2020). Traditional deep learning techniques, particularly convolutional neural networks (CNNs) (Liu et al., 2016) and models like YOLO (You Only Look Once) (Redmon et al., 2016), have been widely utilized to classify individuals in visual data as helmeted or non-helmeted. While these CNN-based models have demonstrated considerable success, they often face challenges in complex scenes characterized by crowded backgrounds, partial occlusions, and the small size of helmets, which make accurate detection difficult.

The recent emergence of Vision Transformers (Dosovitskiy et al., 2020) has opened new avenues in object detection tasks by capturing long-range dependencies and global context within images through self-attention mechanisms. Unlike CNNs, which primarily focus on local features, Vision Transformers model relationships across the entire

image, addressing many limitations of CNNs. This makes them promising candidates for challenging detection tasks, such as helmet monitoring in cluttered industrial environments.

In this project, we advance helmet detection by integrating *Moondream2* (Vikhyat, 2024), a Vision-Language Transformer designed to combine visual and textual information for tasks such as visual question answering (VQA). To the best of our knowledge, no prior work has combined vision-language reasoning with helmet detection. This addition goes beyond basic detection, allowing our system to answer compliance-related questions such as "Are there people without helmets in this image?". By blending visual and linguistic reasoning, *Moondream2* provides deeper insights into safety compliance. Initially, the model encounters challenges in accurately identifying helmets; however, task-specific fine-tuning significantly enhances its detection accuracy, underscoring the model’s adaptability to safety-critical environments.

By leveraging transformers’ capability to interpret both visual and linguistic inputs, this project aims to develop a comprehensive multimodal safety compliance system. Through the *Moondream2* for compliance-related question answering, our approach seeks to demonstrate the effectiveness of transformer-based models in enhancing workplace safety monitoring in complex industrial environments.

Our results demonstrate that fine-tuning the *Moondream2* model on the Hard Hat QA dataset significantly enhances its performance across various safety compliance tasks. The fine-tuned model outperforms the pretrained baseline by a wide margin in metrics such as count accuracy, comparison accuracy, and existence accuracy. Furthermore, the challenging nature of our dataset, which includes cluttered scenes and low-quality images, establishes it as a valuable new benchmark for

evaluating vision-language models in real-world scenarios.

## 2 Related Work

**Object Detection with Convolutional Neural Networks** Object detection has historically been dominated by CNN-based models, such as YOLO (Redmon et al., 2016), which employ layered convolutional structures to identify objects in visual data. These architectures have achieved significant success due to their ability to learn hierarchical features from images. However, they face limitations in complex scenes with crowded backgrounds, partial occlusions, and small object sizes. In such scenarios, the localized receptive fields of CNNs may not capture sufficient contextual information for precise object identification (Liu et al., 2018). These challenges are particularly pronounced in datasets like Hard Hat QA, where real-world complexities dominate.

**Object Detection with Vision Transformers and DETR** Vision Transformers (Dosovitskiy et al., 2020) have recently emerged as a powerful alternative to CNNs for image recognition tasks. By utilizing self-attention mechanisms, Vision Transformers capture global context and model relationships across the entire image. The DETection TRansformer (DETR) (Carion et al., 2020) further advances this field by eliminating the need for region proposal networks, treating images as sequences of patches. While DETR excels in dense, cluttered environments, its focus on pure visual tasks limits its applicability in multimodal scenarios requiring textual reasoning, as highlighted in tasks like those found in the Hard Hat QA dataset.

**Hybrid and Multimodal Models** Hybrid models that combine CNNs and transformers leverage the strengths of both architectures for enhanced performance. *Moondream2* (Vikhyat, 2024), for example, integrates lightweight CNN layers with transformers, achieving computational efficiency while maintaining high-resolution processing. This balance is crucial for tasks like helmet detection in complex industrial environments. Multimodal models, such as LLaVA (Large Language and Vision Assistant) (Liu et al., 2023), go further by incorporating language model capabilities, enabling tasks like visual question answering (VQA) and compliance-related queries. The success of these models underscores the growing importance of integrating visual and textual modalities for real-world

applications. While these models have shown potential, none address helmet detection in real-world industrial settings by combining multimodal reasoning and fine-tuning on task-specific datasets.

**Vision-Language Models and Real-World Benchmarks** The integration of vision and language has expanded the capabilities of Vision Transformers beyond traditional image recognition. Models like *Moondream2* and LLaVA effectively bridge the gap between image understanding and language reasoning, enabling applications such as VQA and compliance-related question answering. However, most existing benchmarks, such as POPE, focus on high-quality images and fail to capture the complexities of real-world industrial environments. The Hard Hat QA dataset, introduced by us, addresses this gap by introducing challenging conditions like low-light scenes, partial occlusions, and imbalanced class distributions, providing a more realistic evaluation of model performance.

**Fine-Tuning and Transfer Learning** Vision Transformers benefit significantly from transfer learning and fine-tuning techniques pioneered in NLP transformers (Devlin et al., 2018). Fine-tuning on domain-specific datasets like Hard Hat QA enables these models to adapt effectively to challenging tasks, such as helmet detection and safety compliance. Recent studies (Kolesnikov et al., 2020) have demonstrated that fine-tuning Vision Transformers can yield substantial improvements in accuracy, particularly for complex datasets. Our work builds on this foundation by demonstrating that fine-tuning *Moondream2* on the Hard Hat QA dataset not only achieves significant performance gains but also underscores the critical importance of adapting vision-language models to the unique complexities of real-world industrial contexts.

## 3 Proposed Methodology

### 3.1 Data Curation and Question-Answer Pairs Generation

The Hard Hat dataset was specifically selected for this study due to its relevance in safety-critical construction scenarios. It contains 5,000 images with approximately 21,000 annotated instances spanning three key classes: helmets, heads, and persons. These annotations include bounding boxes that delineate the spatial extent of objects within images. Unlike general-purpose datasets such as POPE (Li et al., 2023), AMBER (Wang et al., 2024), and

GQA (Hudson and Manning, 2019), which focus on high-quality images and balanced distributions, the Hard Hat dataset introduces real-world complexity. Its defining characteristics include domain-specific application and class imbalance. Images exhibit challenging conditions such as low-light environments, distant viewpoints, and cluttered scenes, where objects are often partially occluded, increasing the difficulty of detection and reasoning tasks. Furthermore, the dataset is imbalanced, with the ‘head’ and ‘person’ classes being significantly underrepresented compared to the ‘helmet’ class (Fig. 1). This presents practical challenges in training models that are robust to real-world data distributions.

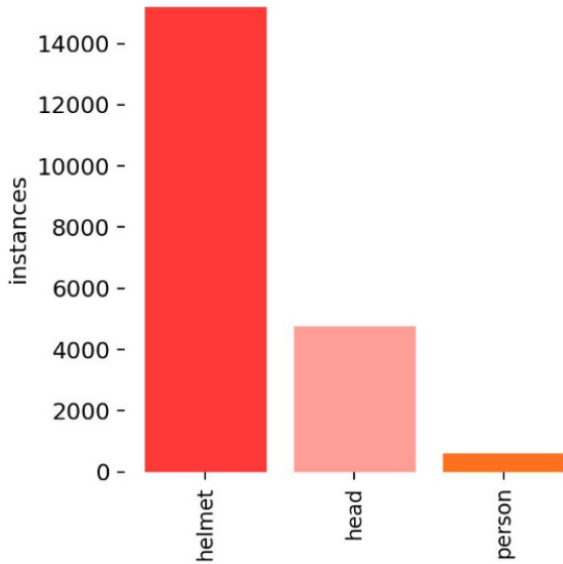


Figure 1: Class distribution in the Hard Hat dataset highlights imbalance across ‘helmet’, ‘head’ and ‘person’ classes.

To adapt the Hard Hat dataset for visual language modeling (VLM), we undertook several preprocessing steps to enhance its usability for multimodal reasoning tasks. The data cleaning process involved validating the bounding box annotations to ensure accuracy and consistency. A subset of images was manually inspected to verify correct labeling and appropriate bounding box coverage. Furthermore, the ‘person’ class was ultimately discarded, simplifying the task into a binary object detection problem: the presence of a helmet or head.

In its original form, the Hard Hat dataset lacked question-answer (QA) pairs, which are essential for enabling multimodal reasoning in VLMs. Inspired by datasets like POPE and AMBER, we

transformed the Hard Hat dataset into a multimodal resource by generating QA pairs from its bounding box annotations. Each image was associated with eight questions spanning three categories:

- **Count-based questions:** We present these pairs in Table 1. These questions require the model to count the number of objects or individuals in the image (Table 1).
- **Existence questions:** These pairs are featured in Table 2. These are binary yes/no questions focusing on whether certain objects or conditions were present.
- **Comparison questions:** These pairs in Table 3. These required models to compare quantities or relationships within the image.

Templates were carefully designed for each question type to align with natural language conventions while avoiding ambiguity. For instance, comparison questions followed the format: "Are there more [Object A] than [Object B]?" Count-based questions used: "How many [Object] are in this image?" An automated script was developed to generate QA pairs by parsing the bounding box annotations. This script calculated object counts, checked for the existence of specific classes, and performed comparisons. For each image, the script generated eight QA pairs spanning all three categories. Manual verification was conducted on a random subset of the generated pairs to ensure accuracy.

The transformed Hard Hat dataset now includes 5,000 images with bounding box annotations and approximately 40,000 QA pairs across the three categories. This enhanced dataset serves as a valuable resource for training and evaluating VLMs in safety-critical domains. By integrating visual and textual reasoning, the dataset enables models to perform quantitative and qualitative analyses of real-world scenarios, bridging the gap between traditional vision tasks and multimodal applications.

### 3.2 Modeling

This study investigates the effectiveness of the *Moondream2* model, a lightweight vision-language transformer, for domain-specific safety compliance tasks. Our methodology focuses on evaluating the pretrained *Moondream2* model and its fine-tuned version on the Hard Hat dataset.

**Pretrained *Moondream2* Model** *Moondream2* is optimized for edge devices with 1.86 billion parameters and achieves competitive performance on







Numerical QA Pairs	
 <i>How many people are in this image?</i>	 (Sum of helmeted and unhelmeted individuals)
 <i>How many hard hats are in this image?</i>	 (Count of helmets)
 <i>How many people without hard hats are in this image?</i>	 (Count of unhelmeted people)

Table 1: Question-based QA pairs generated from the Hard Hat Dataset.







Existence QA Pairs	
 <i>Are there any people in this image?</i>	 Yes / No
 <i>Are all people wearing hard hats in this image?</i>	 Yes / No
 <i>Are there people without hard hats in this image?</i>	 Yes / No

Table 2: Existence QA pairs generated from the Hard Hat Dataset.





Comparison QA Pairs	
 <i>Are there more people with hard hats than without?</i>	 Yes / No (comparing people count and helmet count)
 <i>Are there more people without hard hats than with?</i>	 Yes / No (comparing people count and helmet count)

Table 3: Comparison QA pairs generated from the Hard Hat Dataset.

benchmarks like VQAv2 and POPE. While efficient, its pretrained state is limited in handling domain-specific challenges such as helmet detection in visually complex scenes. This necessitates

fine-tuning to adapt to the specific demands of construction safety applications.

**Fine-Tuning Moondream2** Fine-tuning was performed using the domain-specific QA pairs derived from the Hard Hat dataset. This process enhanced the model’s understanding of safety-critical tasks.

## 4 Experimental Setup

### 4.1 Datasets

We used the QA dataset described previously, which was derived from the Hard Hat dataset. All images were resized to  $512 \times 512$  pixels, and pixel values were normalized to the range  $[0, 1]$ . The dataset was divided into training, validation, and testing sets in a 70:15:15 ratio, and the performance was reported on the test set.

### 4.2 Experimental Procedure and Details

The pretrained Moondream2 model was used as the baseline for comparison. The fine-tuning process aimed to enhance the model’s performance on domain-specific QA tasks by leveraging the enhanced Hard Hat dataset. Both models were evaluated using the generated QA pairs. The model was optimized using the AdamW optimizer with a learning rate of  $10^{-5}$  and a batch size of 16, and training was conducted for 20 epochs on an NVIDIA RTX 3050 GPU.

### 4.3 Evaluation Metrics

The performance of the models was assessed across four key metrics. **Count accuracy** evaluates the model’s ability to correctly quantify objects in an image, such as determining the number of helmets or individuals present. This metric reflects the model’s numerical reasoning capabilities. **Comparison accuracy** measures the capacity to analyze numerical relationships within an image, such as determining whether more individuals are wearing hard hats than not. **Existence accuracy** focuses on binary yes/no questions, such as "Are all individuals wearing hard hats?", providing insight into the model’s scene-specific detection abilities. Finally, textbfmean absolute error (MAE) quantifies the average difference between predicted and actual counts in count-based tasks. Lower MAE values indicate greater precision in numerical predictions, particularly for challenging scenarios.



## 5 Results

### 5.1 Quantitative Results

The quantitative evaluation demonstrates the significant performance improvements achieved by the fine-tuned Moondream2 model compared to its pretrained counterpart across various safety compliance tasks. Table 4 summarizes the aggregated results, while Figure 2 provides a question-level breakdown, offering deeper insights into model performance.

The fine-tuned model exhibits a notable improvement in **count-based questions**, with accuracy increasing from 0.27 to 0.62. This improvement underscores the model’s enhanced numerical reasoning capabilities. Similarly, **comparison-based questions**, which involve evaluating numerical relationships, saw accuracy rise from 0.59 to 0.82, demonstrating the model’s ability to adapt to domain-specific tasks. For **existence-based questions**, the fine-tuned model achieves an accuracy of 0.84, improving upon the pretrained model’s 0.76 and highlighting its consistency in binary decision-making.

Additionally, the fine-tuned model significantly reduces the **mean absolute error (MAE)** for counting tasks, from 1.58 to 0.94, reflecting a substantial improvement in numerical precision. These results are further validated by Figure 2, which illustrates accuracy improvements across all individual question types, particularly for complex queries such as "Are there more people with hard hats than without?" and "How many people are in this image?"

Overall, the quantitative results indicate that fine-tuning the Moondream2 model with domain-specific QA pairs not only improves task-specific performance but also enhances the robustness and reliability of the model in challenging, safety-critical environments.

### 5.2 Qualitative Analysis of Model Predictions

Figures 3, 4, and 5 illustrate the qualitative performance of the pretrained and fine-tuned models on various real-world construction scenes. Each figure highlights the strengths of the fine-tuned model and the limitations of the pretrained baseline model in handling safety compliance tasks.

In Figure 3, the scene contains multiple small, similarly colored hard hats. The fine-tuned model answers all questions correctly and consistently. However, the pretrained model struggles, providing conflicting answers. For instance, it incorrectly

Metric	Pretrained	Fine-Tuned
Count Accuracy	0.27	0.62
Comparison Accuracy	0.59	0.82
Existence Accuracy	0.76	0.84
Mean Absolute Error	1.58	0.94

Table 4: Aggregated performance metrics for pretrained and fine-tuned models across all tasks. The fine-tuned model demonstrates substantial improvements across all metrics.

claims that two people are not wearing hard hats, yet simultaneously states that all individuals are wearing hard hats. This inconsistency underscores the baseline model’s lack of robustness in challenging environments.

Figure 4 features a scene with two different colors of hard hats, adding complexity to the task. The fine-tuned model maintains its consistent and accurate performance, correctly answering all questions. Conversely, the pretrained model continues to provide conflicting answers, mirroring the behavior observed in the previous example.

In Figure 5, the image includes both helmeted and unhelmeted individuals. While the pretrained model fails to accurately predict the number of people and hard hats, the fine-tuned model provides significantly closer predictions, even though it does not perfectly match the ground truth. This highlights the fine-tuned model’s improved accuracy and reliability, particularly for count-based questions in complex and cluttered scenes.

## 6 Discussion and Conclusion

The results of this study highlight the effectiveness of fine-tuning the Moondream2 model using the QA dataset derived from the Hard Hat dataset. Our experimental findings demonstrate that the new dataset provides a robust and challenging benchmark for evaluating vision-language models in safety-critical domains. Unlike existing benchmarks such as POPE, which consist of high-quality and balanced images, our dataset introduces real-world complexity through low-light environments, distant viewpoints, and cluttered scenes. These factors make it uniquely valuable for assessing models in more practical, real-world scenarios.

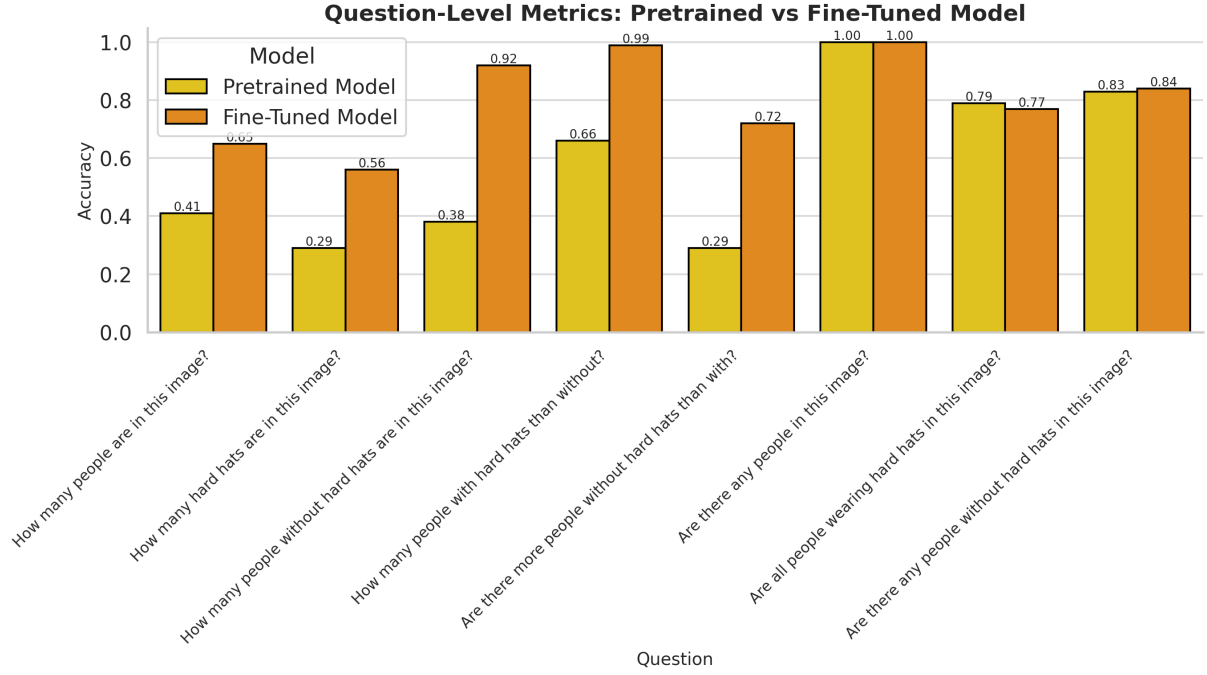


Figure 2: Question-level accuracy comparison between pretrained and fine-tuned models. The fine-tuned model demonstrates significant improvements across all question types, particularly for count-based and comparison-based tasks.

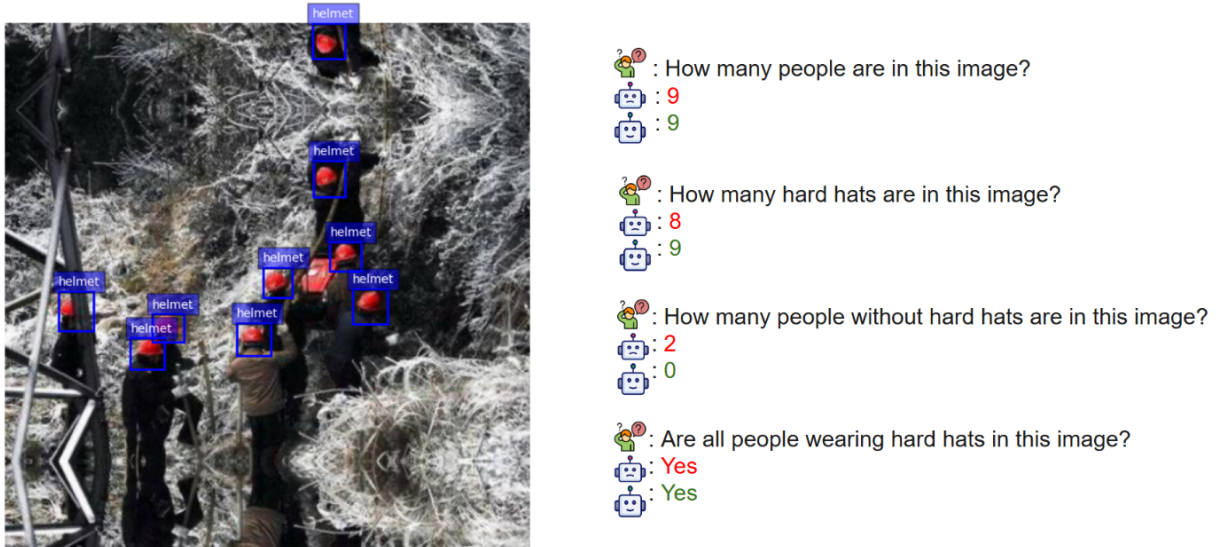


Figure 3: A scene containing multiple small, similarly colored hard hats. The fine-tuned model answers all questions correctly, while the pretrained model provides conflicting and inconsistent answers.

The performance of the pretrained model underscores the difficulty of the dataset, as it consistently fails to produce accurate predictions across count, comparison, and existence-based questions. In contrast, the fine-tuned Moondream2 model significantly improves accuracy and robustness, demonstrating the potential for domain-specific fine-tuning to overcome these challenges. These results establish the Hard Hat QA dataset as a new

benchmark for developing and evaluating models that must perform under complex and imbalanced real-world conditions.

In conclusion, this work provides a foundational step towards improving multimodal reasoning and object detection in safety-critical applications. By introducing a challenging and domain-specific dataset, we have laid the groundwork for further research into improving model architectures



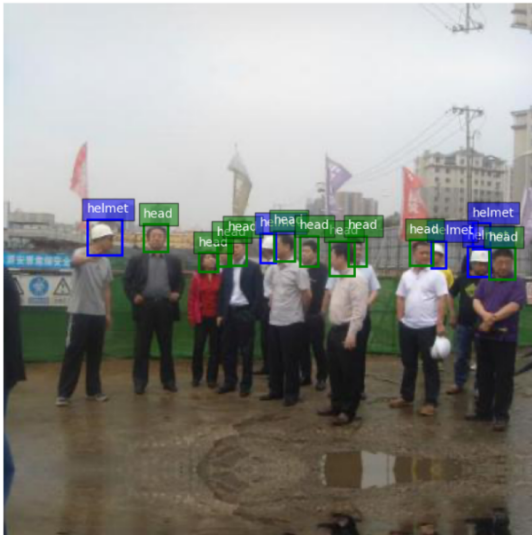
? : How many people are in this image?  
 ? : 6  
 ? : 9

? : How many hard hats are in this image?  
 ? : 8  
 ? : 9

? : How many people without hard hats are in this image?  
 ? : 2  
 ? : 0

? : Are all people wearing hard hats in this image?  
 ? : Yes  
 ? : Yes

Figure 4: A scene with two different colors of hard hats. The fine-tuned model maintains consistent and accurate performance, whereas the pretrained model continues to provide conflicting answers.



? : How many people are in this image?  
 ? : Can't define  
 ? : 14 (instead of 15)

? : How many hard hats are in this image?  
 ? : Can't define  
 ? : 7 (instead of 5)

? : How many people without hard hats are in this image?  
 ? : 4  
 ? : 7 (instead of 10)

? : Are there any people without hard hats?  
 ? : No  
 ? : Yes

Figure 5: A challenging scene with both helmeted and unhelmeted individuals. The fine-tuned model provides significantly closer predictions, while the pretrained model fails to accurately predict counts or compliance.

and training methodologies. Future work could explore augmenting the dataset with additional scenarios, employing transfer learning from other domains, or integrating novel attention mechanisms to further enhance performance on safety compliance tasks. This study demonstrates the importance of designing benchmarks that reflect real-world challenges, ensuring models are equipped to handle the complexities of practical applications.

**Note:** All group members contributed equally to every phase of the project, including dataset preparation, coding, fine-tuning, error analysis, and documentation. The experiments were executed on Eleanna’s lab server, which is why the code is sub-

mitted to the repository under her account.

## References

- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020.

- An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.
- HexArmor. 2019. The hard truth about safety helmet injuries and statistics. <https://www.hexarmor.com/posts/the-hard-truth-about-safety-helmet-injuries-and-helmet-stills>. Accessed: 2024-11-04.
- Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.
- Sung Hun Kim et al. 2018. Safety helmet wearing management system for construction workers using three-axis accelerometer sensor. *Applied Sciences*, 8(12):2400.
- Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision (ECCV)*, pages 491–507.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Zongyu Li and Min Li. 2020. Real-time helmet detection for construction safety using deep learning. *Journal of Electrical and Computer Engineering*, 2020:1–8.
- Jie Liu, Zhenfang Ye, Xudong Zou, Zhaoyang Liu, Hong Diao, Xiaodan Liang, Liang Zhang, and Xiaojun Hu. 2023. Llava: Large language and vision assistant for unified vision-language tasks. *arXiv preprint arXiv:2303.17206*.
- Shifeng Liu, Di Huang, and Yunhong Wang. 2018. Receptive field block net for accurate and fast object detection. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 404–419.
- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37.
- S Mahalakshmi, K Meena, and N Rajeswari. 2020. Helmet detection system using machine learning and image processing techniques. *Materials Today: Proceedings*, 37:1850–1856.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.
- Vikhyat. 2024. Moondream2: A tiny vision-language model. <https://huggingface.co/vikhyatk/moondream2>. Accessed: 2024-11-03.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024. Amber: An unified multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.