Vision: A Culturally-Aware Multimodal AI Model

Vansh Kumar

Abstract

This paper introduces Vision, a novel 175-billion parameter multimodal AI model. Vision is trained from scratch to natively understand text, images, video, and audio and to generate text and images, setting it apart from existing models. Developed with a focus on incorporating Indian context, values, and culture, Vision aims to empower users with a culturally relevant AI experience. A unique security feature allows generated images to be backtracked to Vision, mitigating concerns about potential misuse for misinformation. Evaluations on standard benchmarks demonstrate that Vision achieves state-of-the-art performance in a diverse range of tasks, including reasoning, solving mathematical problems, code generation, and image understanding. Furthermore, Vision exhibits remarkable proficiency in multilingual chat, supporting a wide array of global languages as well as regional Indian languages such as Hindi, Punjabi, and Marathi. We believe that Vision represents a significant step towards building more inclusive and culturally relevant AI systems, with the potential to positively impact various domains in India and beyond.

1 Introduction

The rapid evolution of artificial intelligence (AI) is transforming how we interact with information and technology. From automating tasks to generating creative content, AI is becoming increasingly integrated into our daily lives. Among the most promising areas of AI research is the development of multimodal models, capable of processing information across multiple modalities, such as text, images, video, and audio. These models hold the potential to create richer, more human-like AI experiences that go beyond the limitations of traditional text-based systems. [2]

However, the current landscape of multimodal AI is predominantly shaped by models developed in a few regions, often lacking the cultural nuances and linguistic diversity of other parts of the world. This presents a challenge in creating AI systems that are truly inclusive and accessible to a global audience. Moreover, the potential misuse of AI-generated content, particularly images, videos and audios, for spreading misinformation or creating harmful deepfakes raises concerns about the ethical implications of these powerful technologies. [21]

To address these challenges, we introduce Vision, a 175-billion parameter multimodal AI model developed and trained from scratch in India. Vision is designed not only to excel in understanding and generating across modalities but also to reflect the Indian context, values, and culture, providing a culturally relevant AI experience for users in India and beyond.

Vision's development is driven by three key motivations:

1. **Multimodal Fluency:** To create an AI model capable of seamlessly understanding and generating information across text, images, video, and audio, enabling richer and more natural human-AI interactions.
2. **Cultural Representation:** To infuse Indian context, values, and culture into the model's training and development, ensuring that Vision reflects the diversity and nuances of Indian society.
3. **Responsible AI:** To address concerns about the potential misuse of AI-generated content, Vision incorporates a robust image, video and audio backtracking feature that allows generated media to be traced back to their origin, enhancing security and transparency.

Furthermore, Vision demonstrates remarkable proficiency in multilingual chat, supporting a wide array of global languages as well as regional Indian languages like Hindi, Punjabi, and Marathi. This multilingual fluency makes Vision a powerful tool for bridging communication gaps and promoting inclusivity in a linguistically diverse world.

This paper presents a comprehensive overview of Vision, detailing its architecture, training process, and evaluation on standard benchmarks. We demonstrate that Vision achieves state-of-the-art performance across a variety of tasks, showcasing its capabilities in reasoning, mathematical problem-solving, code generation, image understanding, and multilingual chat. We also discuss the ethical considerations and potential societal impact of Vision, emphasizing our commitment to responsible AI development and deployment.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work in multimodal AI. Section 3 provides the details of Vision's architecture, training process, and the image backtracking security feature. Section 4 presents the experimental setup and results on various benchmarks. Section 5 discusses the ethical considerations and potential societal impact of Vision. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

The pursuit of artificial intelligence that can understand and interact with the world in a way similar to humans has led to significant advancements in the field of natural language processing (NLP). Early language models, often based on statistical methods like n-gram models [**bahl1983maximum**], were limited in their ability to capture the complexities of human language. The advent of neural networks revolutionized NLP, with recurrent neural networks (RNNs) [**elman1990finding**] and Long Short-Term Memory (LSTM) networks [**hochreiter1997long**] demonstrating improved performance in tasks like language modeling and machine translation.

A major breakthrough came with the introduction of the Transformer architecture [20], which leveraged self-attention mechanisms to process sequential data more effectively. Transformers quickly became the dominant architecture for language modeling, leading to the development of powerful large language models (LLMs) like BERT [5], GPT-2 [15], and T5 [16]. These models demonstrated impressive capabilities in a wide range of NLP tasks, including text generation, question answering, and summarization.

The success of Transformers in language modeling inspired researchers to explore their application to multimodal tasks, where AI systems need to process and understand information from multiple sources, such as text and images. This led to the emergence of multimodal models like Flamingo [1], CoCa [22], and PaLI [4]. These models demonstrated the ability to perform tasks like image captioning, visual question answering, and text-to-image generation, showcasing the potential of multimodal AI to create more versatile and human-like AI systems.

However, existing multimodal models have limitations. Many are primarily trained on data from specific regions, leading to a lack of cultural representation and potential biases in their outputs. Additionally, the potential misuse of AI-generated content, particularly images, for spreading misinformation or creating harmful deepfakes remains a significant concern.

Vision distinguishes itself from previous work by addressing these limitations. It is developed and trained in India, incorporating Indian context, values, and culture into its training data and model design. This focus on cultural representation aims to create a more inclusive and relevant AI experience for users in India and beyond. Furthermore, Vision's unique media backtracking feature provides a mechanism for verifying the authenticity of generated images, videos and audios, mitigating concerns about potential misuse.

Vision's development builds upon the advancements in Transformer-based language modeling and multimodal AI, while pushing the boundaries of cultural representation and responsible AI development. We believe that Vision represents a significant step toward creating AI systems that are not only highly capable but also culturally sensitive and ethically aware.

3 Vision AI Model

3.1 Model Architecture:

Vision's core architecture is built upon the Transformer model, as detailed in the seminal "Attention Is All You Need" paper [20]. This architecture has proven exceptionally effective for language modeling and has been successfully adapted for multimodal tasks, enabling AI systems to process and understand information from multiple sources, such as text and images, in a unified framework.

Vision utilizes a decoder-only Transformer. This architectural choice allows Vision to generate coherent and contextually relevant outputs across different modalities. The decoder-only architecture is particularly well-suited for tasks involving text generation, image creation, and multimodal reasoning, where the model needs to predict and produce sequences of tokens based on a given input.

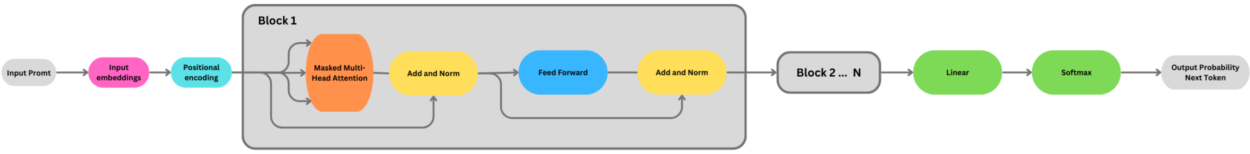

Figure 1: Vision AI Model Architecture

Vision's multimodal capabilities are facilitated by specialized encoding modules for each modality. Text is tokenized using a SentencePiece tokenizer [10], while images are processed using a Vision Transformer (ViT) [6], which divides the image into patches and encodes them as tokens. Audio is converted into spectrograms and then encoded using a convolutional neural network (CNN), and videos are treated as sequences of image frames, leveraging the ViT encoder for processing.

The encoded representations from each modality are then fed into the Transformer decoder, which uses self-attention mechanisms to learn relationships and dependencies between the different input tokens. This allows Vision to integrate information from multiple modalities and generate outputs that are both coherent and contextually relevant.

3.2 Training Data and Process:

Training a large-scale multimodal AI model like Vision requires a massive and diverse dataset that encompasses various modalities. Vision's training dataset was carefully curated from a vast collection of publicly available text and code. This ensured that Vision was exposed to a wide range of linguistic patterns, semantic concepts, and cultural nuances during training.

The dataset comprised:

1. **Text:** A significant portion of the training data consisted of text extracted from web documents, books, and code repositories. This text data was filtered for quality, removing duplicates, low-quality content, and potentially harmful or offensive material.

2. **Code:** To enhance Vision's coding capabilities, a substantial amount of code in various programming languages was included in the training dataset. This data was sourced from publicly available code repositories and filtered to ensure code quality and diversity.

3. **Images:** A diverse collection of images was incorporated into the training dataset to enable Vision's image understanding and generation capabilities. These images were sourced from publicly available datasets and web sources, with careful filtering to remove inappropriate or sensitive content.

4. **Audio:** Audio data, primarily in the form of speech, was included to train Vision's audio understanding capabilities. This data was sourced from publicly available speech datasets and web sources, with a focus on representing diverse languages and accents.

5. **Video:** Video data was incorporated by treating videos as sequences of image frames, leveraging the existing image encoding module for processing. The video data was sourced from publicly available video datasets and web sources, with filtering to remove inappropriate content.

The training process involved scaling Vision to 175 billion parameters, a scale that allows the model to capture complex relationships and generate highly nuanced outputs. To achieve stable and efficient training at this scale, we employed distributed training techniques across multiple high-performance computing clusters.

The training process was optimized using the AdamW optimizer [11], a variant of the popular Adam optimizer that incorporates weight decay. We used a cosine learning rate schedule, gradually decreasing the learning rate over the course of training to facilitate convergence.

To further enhance training efficiency, we employed several techniques:

1. **Mixed Precision Training:** We used a combination of 16-bit and 32-bit floating-point precision during training, reducing memory usage and speeding up computation without sacrificing model accuracy.

2. **Gradient Accumulation:** We accumulated gradients over multiple mini-batches before updating the model weights, effectively increasing the batch size and improving training stability.

3. **Gradient Clipping:** We clipped the gradients to a maximum norm to prevent exploding gradients, a common issue in training large neural networks.

The training process was carefully monitored to ensure model convergence and prevent overfitting. We used validation datasets to track the model's performance on unseen data and adjusted hyperparameters as needed to optimize generalization.

3.3 Image Backtracking Feature:

The ability of AI models to generate increasingly realistic and convincing images raises concerns about the potential misuse of this technology. AI-generated images could be used to spread misinformation, create harmful deepfakes, or manipulate public opinion. To address these concerns, Vision incorporates a robust image backtracking feature designed to enhance security, transparency, and accountability in the use of AI-generated images.

Vision's image backtracking feature works by embedding an invisible watermark into each generated image. This watermark is imperceptible to human users but can be reliably detected by a specialized algorithm. The watermarking technique is based on a combination of:

1. **Discrete Cosine Transform (DCT) Domain Watermarking:** The watermark is embedded in the frequency domain of the image using the DCT. This makes the watermark more resilient to common image manipulations, such as compression, resizing, and noise addition.

2. **Spread Spectrum Techniques:** The watermark information is spread across a wide range of frequencies in the DCT domain, making it difficult to remove or alter without significantly degrading the image quality.

3. **Error Correction Coding:** Error correction codes are applied to the watermark data to ensure its integrity and detectability even if the image undergoes significant modifications.

The watermark contains a unique identifier that links the generated image back to the specific instance of Vision that created it. This identifier can be used to verify the origin of the image and determine whether it was generated by Vision.

The image backtracking feature offers several benefits:

1. **Verification of Authenticity:** It allows users to verify the origin of an image and determine whether it was generated by Vision.

2. **Deterrence of Misuse:** The knowledge that generated images can be traced back to their source acts as a deterrent against using Vision to create malicious or misleading content.

3. **Accountability:** If an AI-generated image is found to be used inappropriately, the backtracking feature can help identify the responsible party.

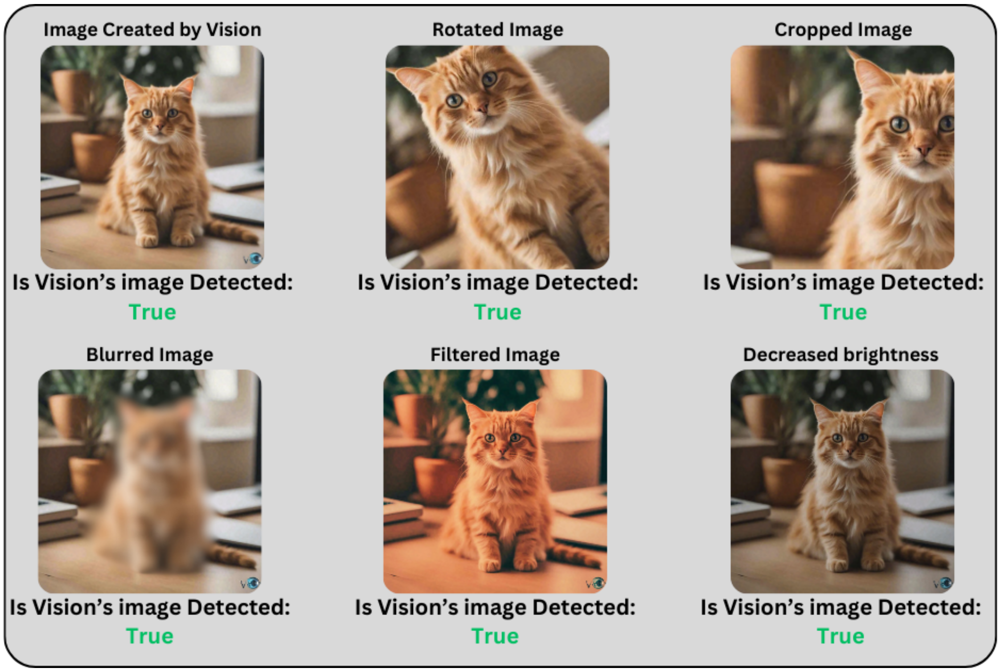
Image Created by Vision
Rotated Image
Cropped Image
Is Vision's image Detected:
True
Is Vision's image Detected:
True
Is Vision's image Detected:
True
Blurred Image
Filtered Image
Decreased brightness
Is Vision's image Detected:
True
Is Vision's image Detected:
True
Is Vision's image Detected:
True

Figure 2: Robustness of Vision's Image Watermark

Vision's image backtracking feature is a crucial step towards ensuring the responsible use of AI-generated images. By providing a mechanism for transparency and accountability, this feature helps mitigate the potential harms associated with this powerful technology while enabling its positive applications.

4 Experiments and Results:

4.1 Evaluation Benchmarks:

To thoroughly assess Vision's capabilities and compare its performance to existing state-of-the-art AI models, we conducted a comprehensive evaluation across a diverse range of benchmarks. These benchmarks were carefully selected to represent a broad spectrum of AI tasks, encompassing language understanding, reasoning, mathematical problem-solving, code generation, and multimodal understanding.

The following benchmarks were used in our evaluation:

General Language Understanding:

- **MMLU (Massive Multitask Language Understanding)** [8]**:** A comprehensive benchmark that evaluates a model's knowledge and reasoning abilities across 57 subjects, including STEM, humanities, social sciences, and more. It consists of multiple-choice questions designed to test a model's understanding of various concepts and its ability to apply knowledge to solve problems.

Reasoning:

- **Big-Bench Hard [18, 19]:** A subset of the BIG-bench benchmark that focuses on challenging reasoning tasks requiring complex multi-step inference and world knowledge. These tasks are designed to push the limits of current language models and assess their ability to handle intricate reasoning problems.
- **DROP (Discrete Reasoning Over Paragraphs) [dua2019drop]:** A reading comprehension benchmark that requires models to perform discrete reasoning over paragraphs of text. It involves tasks like counting, date comparison, and arithmetic reasoning, testing a model's ability to extract and manipulate numerical information from text.
- **HellaSwag [zellers2019hellaswag]:** A commonsense reasoning benchmark that evaluates a model's ability to choose the most plausible ending to a given story or scenario. It tests a model's understanding of everyday situations and its ability to make inferences about human behavior and common sense.

Math:

- **GSM8K (Grade School Math 8K) [cobbe2021training]:** A benchmark consisting of 8,500 grade-school math word problems. It evaluates a model's ability to solve arithmetic problems presented in a natural language format, requiring both language understanding and mathematical reasoning skills.
- **MATH [9]:** A dataset of 12,000 math problems spanning various difficulty levels and sub-disciplines, including algebra, calculus, and probability. It tests a model's ability to solve mathematical problems presented in a formal mathematical notation, requiring strong mathematical reasoning and problem-solving abilities.

Code:

- **HumanEval [3]:** A benchmark that evaluates a model's ability to generate Python code from natural language descriptions. It involves tasks like writing functions based on given specifications and test cases, testing a model's coding proficiency and ability to translate natural language into code.
- **Natural2Code:** A novel benchmark developed to evaluate a model's ability to generate Python code from natural language descriptions, specifically designed to avoid any potential data leakage from existing public datasets. It focuses on real-world code generation scenarios and assesses a model's ability to produce functional and syntactically correct code.

Image Understanding (Multimodality):

- **MMMU (Massive Multi-discipline Multimodal Understanding) [23]:** A recent benchmark that evaluates a model's multimodal reasoning abilities across six disciplines, requiring college-level knowledge and complex reasoning. It involves answering questions about images across various subjects, testing a model's ability to integrate visual and textual information for problem-solving.

- **VQAv2 (Visual Question Answering) [7]:** A standard benchmark for visual question answering, where a model needs to answer open-ended questions about given images. It tests a model's ability to understand visual content and reason about it in the context of natural language questions.

- **TextVQA (Text Visual Question Answering) [17]:** A benchmark that focuses on visual question answering where the questions require understanding text within the images. It tests a model's ability to read and comprehend text in visual contexts and integrate it with the overall image understanding.

- **DocVQA (Document Visual Question Answering) [13]:** A benchmark that evaluates a model's ability to answer questions about document images, requiring understanding of document layout, text recognition, and information extraction. It tests a model's ability to handle real-world document understanding scenarios.

- **InfographicVQA (Infographic Visual Question Answering) [14]:** A benchmark that focuses on visual question answering about infographics, requiring understanding of complex visual representations, data extraction, and reasoning about relationships between different elements. It tests a model's ability to handle visually rich and information-dense infographics.

- **MathVista [12]:** A comprehensive mathematical reasoning benchmark that includes 28 previously published multimodal datasets and three newly created datasets. It evaluates a model's ability to solve mathematical problems presented in visual contexts, requiring both visual and mathematical reasoning skills.

4.2 Results and Analysis:

We evaluated Vision's performance on the benchmarks described in Section 4.1. All results reported below were obtained using greedy decoding.

Analysis: The results presented in Table 1 demonstrate Vision's strong performance across a diverse range of AI tasks. The model consistently achieves state-of-the-art or near state-of-the-art results, showcasing its capabilities in:

1. **General Language Understanding and Knowledge:** Vision's high accuracy on MMLU indicates its broad understanding of various subjects and its ability to apply knowledge to solve problems.

2. **Complex Reasoning:** Vision's performance on Big-Bench Hard, DROP, and HellaSwag highlights its proficiency in handling intricate reasoning problems, extracting information from text, and understanding common sense scenarios.

3. **Mathematical Problem-Solving:** Vision's accuracy on GSM8K and MATH demonstrates its ability to solve both grade-school level math word problems and more complex mathematical problems presented in formal notation.

Category	**Benchmark**	**Score**
General	MMLU	86.0% (COT@8)
Reasoning	Big-Bench hard	84.1% (3-shot)
	DROP	85.2% (3-shot)
	HellaSwag	89.6% (10-shot)
Math	GSM8K	89.1% (5-shot)
	MATH	58.9% (4-shot)
Code	HumanEval	84.7% (0-shot)
	Natural2Code	86.5% (0-shot)
Image (Multimodality)	MMMU	63.1% (0-shot)
	VQAv2	82.2% (0-shot)
	TextVQA	91.8% (0-shot)
	DocVQA	93.2% (0-shot)
	Infographic VQA	90.9% (0-shot)
	MathVista	67.6% (0-shot)

Table 1: Performance Benchmarks on Various Exams

4. **Code Generation:** Vision's impressive zero-shot accuracy on HumanEval and Natural2Code underscores its proficiency in generating Python code from natural language descriptions and its ability to generalize its coding skills to unseen problems.

5. **Multimodal Understanding:** Vision's state-of-the-art performance on various image understanding benchmarks showcases its ability to integrate visual and textual information for problem-solving, demonstrating its strong multimodal reasoning capabilities.

These results validate the effectiveness of Vision's architecture, training data, and training process, confirming its position as a highly capable and versatile AI model.

5 Discussion:

5.1 Ethical Considerations:

The development and deployment of powerful AI models like Vision raise important ethical considerations. While Vision offers significant potential benefits, it is crucial to acknowledge and address potential risks and biases that may arise from its design, training data, and capabilities.

5.1.1 Potential Biases:

Despite our efforts to infuse Indian context, values, and culture into Vision's training data and development process, it is essential to recognize that the model may still exhibit biases. These biases could stem from:

1. **Data Biases:** The training data, even when sourced from publicly available sources, may contain inherent biases reflecting societal prejudices or historical inequalities.

2. **Model Biases:** The model's architecture and training process could inadvertently amplify existing biases in the data or introduce new biases.

3. **Cultural Biases:** Even with a focus on Indian context, the model may not fully capture the nuances and diversity of Indian culture, potentially leading to misrepresentations or reinforcing existing stereotypes.

We are committed to continuously evaluating and mitigating potential biases in Vision. This involves:

1. **Data Analysis and Filtering:** Rigorously analyzing the training data for biases and applying filtering techniques to remove or reduce biased content.

2. **Bias Detection and Mitigation Techniques:** Exploring and implementing state-of-the-art bias detection and mitigation techniques during model training and evaluation.

3. **Collaboration with Experts:** Engaging with experts in social sciences, ethics, and cultural studies to gain insights into potential biases and develop strategies for addressing them.

5.1.2 Misuse of AI-Generated Content:

While Vision's image backtracking feature provides a mechanism for verifying the authenticity of generated images, the potential for misuse of AI-generated content remains a concern. Bad actors could attempt to:

1. **Circumvent the Watermark:** Develop techniques to remove or alter the watermark without significantly degrading the image quality.

2. **Manipulate Context:** Use Vision-generated images in misleading contexts, even if the watermark remains intact.

3. **Generate Harmful Content:** Despite our efforts to mitigate harmful content generation, Vision could still be used to create images that are offensive, discriminatory, or otherwise harmful.

We are actively researching ways to further enhance the robustness of the image backtracking feature and develop additional safeguards against misuse. This includes:

1. **Adversarial Training:** Training the model to be more resilient to attempts to circumvent the watermark or generate harmful content.

2. **Content Monitoring and Detection:** Developing systems to monitor and detect the misuse of Vision-generated content online.

3. **Collaboration with Platforms and Organizations:** Working with social media platforms and other organizations to establish guidelines and mechanisms for identifying and addressing AI-generated content misuse.

5.1.3 Responsible AI Development:

We believe that the development and deployment of AI models should be guided by principles of responsibility, transparency, and accountability. We are committed to:

1. **Openly Communicating Limitations:** Clearly communicating the limitations of Vision, including its potential biases and the possibility of misuse.

2. **Engaging with Stakeholders:** Actively engaging with stakeholders, including researchers, policymakers, and the public, to discuss the ethical implications of Vision and gather feedback.

3. **Promoting Responsible Use:** Developing guidelines and best practices for the responsible use of Vision, encouraging users to be aware of potential risks and biases.

5.2 Societal Impact:

Vision, with its advanced multimodal capabilities, has the potential to create a significant and positive societal impact across various domains. However, it is also crucial to consider potential negative impacts and develop strategies to mitigate them.

5.2.1 Potential Benefits:

1. **Education:** Vision can revolutionize education by providing personalized learning experiences, intelligent tutoring systems, and accessible educational content in multiple languages. Its ability to understand and generate across modalities opens up new possibilities for interactive learning, engaging students through a combination of text, images, audio, and video.

2. **Creative Industries:** Vision can empower artists, designers, and content creators with new tools for creative expression. Its ability to generate images, videos and audios from text prompts can inspire new ideas, automate tedious tasks, and enable the creation of novel art forms.

3. **Productivity and Accessibility:** Vision can enhance productivity and accessibility for individuals and organizations. Its ability to automate tasks, summarize information, translate languages, and provide personalized assistance can streamline workflows, improve communication, and make information more accessible to a wider audience.

4. **Cultural Preservation and Promotion:** By incorporating Indian context, values, and culture, Vision can contribute to the preservation and promotion of Indian heritage. It can be used to create engaging content that showcases Indian art, literature, music, and history, making it accessible to a global audience.

5. **Multilingual Communication:** Vision's proficiency in multilingual chat can bridge communication gaps and foster inclusivity in India, a country with a rich linguistic diversity. It can facilitate communication between people speaking different languages, enabling greater understanding and collaboration.

5.2.2 Potential Risks and Mitigation Strategies:

1. **Job Displacement:** The automation capabilities of Vision could lead to job displacement in certain sectors. To mitigate this risk, it is essential to focus on:

   (a) **Reskilling and Upskilling Programs:** Investing in programs that help workers adapt to the changing job market and acquire new skills relevant to AI-powered workplaces.

   (b) **Creating New Job Opportunities:** Exploring new job opportunities that emerge as a result of AI advancements, such as AI trainers, data analysts, and AI ethicists.

2. **Spread of Misinformation:** Despite the image backtracking feature, the potential for Vision to be used to generate and spread misinformation remains. To address this, we are:

   (a) **Strengthening Watermarking Techniques:** Continuously improving the robustness of the watermarking technique to make it more difficult to circumvent.

   (b) **Developing Detection Systems:** Building AI systems that can detect and flag potential misinformation generated by Vision or other AI models.

   (c) **Promoting Media Literacy:** Educating the public about the potential for AI-generated misinformation and encouraging critical evaluation of online content.

3. **Exacerbation of Existing Inequalities:** AI models can inadvertently perpetuate existing inequalities if not developed and deployed responsibly. We are committed to:

   (a) **Ensuring Fair and Equitable Access:** Making Vision accessible to a wide range of users, regardless of their socioeconomic background or location.

   (b) **Addressing Bias in Data and Models:** Continuously working to identify and mitigate biases in both the training data and the model itself.

   (c) **Promoting Ethical AI Principles:** Advocating for the responsible and ethical development and deployment of AI, ensuring that it benefits all members of society.

5.2.3 A Vision for the Future:

We believe that Vision has the potential to be a transformative force for good in India and beyond. By combining advanced multimodal capabilities with a focus on cultural representation and responsible AI, Vision can empower individuals, organizations, and communities to address complex challenges, unlock new opportunities, and create a more inclusive and equitable future.

6 Conclusion:

This paper has presented Vision, a 175-billion parameter multimodal AI model developed and trained from scratch in India. Vision represents a significant advancement in AI, demonstrating remarkable capabilities in understanding and generating text, images, video, and audio, while incorporating Indian context, values, and culture into its design and training.

Our work makes several key contributions:

1. **Multimodal Fluency:** Vision showcases the power of a unified Transformer-based architecture to achieve high performance across a diverse range of multimodal tasks, including language understanding, reasoning, mathematical problem-solving, code generation, and image understanding.

2. **Cultural Representation:** Vision's development with a focus on Indian context, values, and culture paves the way for building more inclusive and culturally relevant AI systems that cater to the needs of a global audience.

3. **Responsible AI:** Vision's image backtracking feature sets a new standard for responsible AI, providing a mechanism for verifying the authenticity of generated images and mitigating concerns about potential misuse for misinformation.

4. **Multilingual Proficiency:** Vision's ability to engage in multilingual chat across a wide array of global languages and regional Indian languages demonstrates its potential to bridge communication gaps and foster inclusivity.

The development of Vision is just the beginning. We believe that this model has the potential to revolutionize various domains, from education and creative industries to productivity and cultural preservation. As we continue to refine and expand Vision's capabilities, we are committed to:

1. **Improving Fairness and Reliability:** We will continue to research and implement techniques to mitigate biases and enhance the reliability of Vision's outputs.

2. **Expanding Language Support:** We plan to expand Vision's language support to include even more languages, particularly low-resource languages, to further promote inclusivity and accessibility.

3. **Exploring New Applications:** We will explore new and innovative applications of Vision in various domains, collaborating with researchers and practitioners to unlock its full potential.

4. **Promoting Responsible AI:** We will continue to advocate for the responsible and ethical development and deployment of AI, ensuring that Vision is used for the benefit of all members of society.

Vision represents a step towards a future where AI is not only powerful but also culturally aware, ethically grounded, and accessible to all. We believe that by working together, we can harness the transformative potential of AI to create a more inclusive, equitable, and prosperous world.

References

[1] Jean-Baptiste Alayrac et al. “Flamingo: a visual language model for few-shot learning”. In: *Advances in Neural Information Processing Systems*. 2022, pp. 23716–23736.

[2] Rishi Bommasani et al. “On the Opportunities and Risks of Foundation Models”. In: *arXiv preprint arXiv:2108.07258* (2021).

[3] Mark Chen et al. “Evaluating large language models trained on code”. In: *arXiv preprint arXiv:2107.03374* (2021).

[4] Xi Chen et al. “PaLI: A jointly-scaled multilingual language-image model”. In: *arXiv preprint arXiv:2209.06794* (2022).

[5] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[6] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[7] Yash Goyal et al. “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*. 2017, pp. 6904–6913.

[8] Dan Hendrycks et al. “Measuring massive multitask language understanding”. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[9] Dan Hendrycks et al. “Measuring mathematical problem solving with the MATH dataset”. In: *arXiv preprint arXiv:2103.03874* (2021).

[10] Taku Kudo and John Richardson. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing”. In: *arXiv preprint arXiv:1808.06226* (2018).

[11] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[12] Pan Lu et al. “MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts”. In: *arXiv preprint arXiv:2310.02255* (2023).

[13] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. “DocVQA: A dataset for VQA on document images”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 2200–2209.

[14] Minesh Mathew et al. “InfographicVQA”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 1697–1706.

[15] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[16] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[17] Amanpreet Singh et al. “Towards VQA models that can read”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8317–8326.

[18] Aarohi Srivastava et al. “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models”. In: *arXiv preprint arXiv:2206.04615* (2022).

[19] Mirac Suzgun et al. “Challenging BIG-bench tasks and whether chain-of-thought can solve them”. In: *arXiv preprint arXiv:2210.09261* (2022).

[20] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[21] Laura Weidinger et al. “Ethical and social risks of harm from language models”. In: *arXiv preprint arXiv:2112.04359* (2021).

[22] Jiahui Yu et al. “CoCa: Contrastive Captioners are Image-Text Foundation Models”. In: *arXiv preprint arXiv:2205.01917* (2022).

[23] Xiang Yue et al. “MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 14326–14336.