Evaluating Stable Diffusion's Capability in Generating Context-Appropriate Emotions

A SYSTEMATIC ANALYSIS OF FEAR AND ANGER DEPICTION USING EMOTIONBENCH SCENARIOS

1 Introduction

Generative AI

- **Innovation Pace:** AI innovation in a swift pace
- **Generative AI Adoption:** Widely used in virtual communication and entertainment
  - Virtual chatbots, AI assistants
  - Contents Creation & AI within gaming

Emotions

- **Role of Emotions:** Essential to human interaction
- **Emotional Intelligence:** Would improve user experience and engagement.
- **Alignment studies:** Textual models perform better than image models [1]
- **Model Limitations:** Stable Diffusion struggles to convey intended emotions [2]

Emotions & Context

- **Understanding context:** Crucial for accurate emotional interpretation
- **Context Interpretation:** Stable Diffusion's capabilities are unclear

2 Research

Research Questions

1. **How accurately are fear and anger generated?**
2. **Are there observable biases in the expressions?**
3. **Does prompt specificity influence the accuracy?**

Contributions

- **Filling the Gap:** Explore the challenge of generating emotions from contextual information.
- **Evaluation Framework:** A systematic method to assess emotional accuracy.
- **Guiding Future Research:** Provide insights and recommendations for improving emotional alignment.

Contact Info

J. den boer
J.denboer-3@student.tudelft.nl

Supervisor: Anna Lukina

3 Methods

We created a three-step approach:

1. **Prompt Design:** Prompts crafted from EmotionBench [3] dataset using GPT-4
2. **Image Generation:** Stable Diffusion used to generate images
3. **Emotion Evaluation:** GPT-4V employed to evaluate the generated images

Evaluation Process

1. Take scenario from EmotionBench:

"You get home from the drive-thru and realize that you were given the wrong food." Label: Anger

↓

2. Create three prompts with different amounts of context

↓

3. Generate images using Stable Diffusion:

↓

4. Classify into the following categories using GPT-4V :

(1) Anger, (2) Fear, (3) Other Negative Emotion, (4) Other Positive Emotion, (5) No emotion visible, (6) No person visible

↓

5. Repeat for all anger and fear scenario's that are to be evaluated

4 Results

RQ1 and RQ2: Accuracy of Emotions and Prompt Specificity:

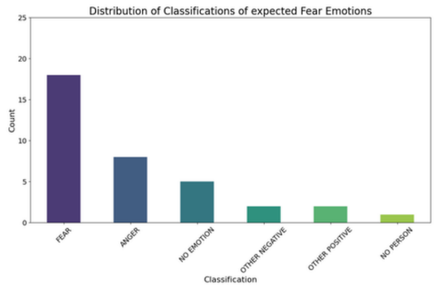
Detail Level	Anger Accuracy (%)	Fear Accuracy (%)
Vague	20.0	45.45
Moderate	9.09	70.0
High	9.09	66.67

Emotion accuracy and impact of prompt specificity.

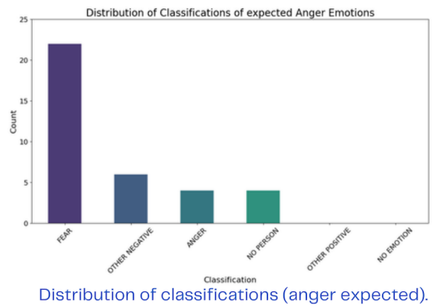
Key Findings:

- Fear overall highest accuracy for moderate detail (70.0%)
- Anger does not follow same trend, highest for vague (20.0%)

RQ3: Biases in emotions



Distrubition of classifications (fear expected).



Distribution of classifications (anger expected).

Key Findings:

- Fear is predominantly correctly classified
- Anger scenario's often misclassified as fear

References

[1] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, and M. R. Lyu, "Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench," Apr. 2024.
[2] J. Lomas, W. van der Maden, S. Bandyopadhyay, G. Lion, Y. Litowsky, H. Xue, P. Desmet, D. Lomas, Yanna, H. Litowsky, and Xue, The Alignment of AI Emotions: human ratings of the emotions expressed by GPT-3, DALL-E and Stable Diffusion, Apr. 2023.
[3] CUHK-ARISE, "Emotionbench/situations at main," https://github.com/CUHK-ARISE/EmotionBench/tree/main/situations, accessed: May 22, 2024

5 Discussion

Model Performance

- **Accuracy and bias**
  - Disparity between fear and anger
  - Bias towards generating fear
  - Possible imbalances in the training data
  - Some emotions may be underrepresented or depicted less accurately
- **Prompt Specificity**
  - Moderate detail prompts work best for fear (70%), but not for anger
  - Variability in results does not support a clear conclusion

Implications

- **Reliance on GPT-4V:** May limit the assessment scope
- **No Defined Specificity:** Might affect result consistency
- **Limited Emotion Range:** Constrained understanding of the model's capabilities
- **Limited Amount Scenarios:** May introduce sampling bias, potentially skewing findings

6 Conclusions & Future Work

Conclusions

- **Fear scenario's:** reasonable accuracy
- **Anger scenario's:** significant struggles
- **Fear-related expressions:** may be better represented in training data
- **One-size-fits-all approach:** might not be effective for prompt specificity

Future Work

- Evaluate a wider range of emotions
- Involve human raters for deeper insights
- Develop consistent methods for defining prompt specificity
- Comparative Analysis of other text-to-image models