Bridging the Emotional Gap: Evaluating Stable Diffusion's Capability in Generating Context-Appropriate Emotions

A Systematic Analysis of Fear and Anger Depiction Using EmotionBench Scenarios

Joosep den Boer[1]
Supervisor(s): Anna Lukina[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Joosep den Boer
Final project course: CSE3000 Research Project
Thesis committee: Anna Lukina, Petr Kellnhofer

Abstract

The ability of generative AI models to accurately depict emotional expressions is crucial for their use in virtual communication and entertainment. This study evaluates Stable Diffusion's capability to generate context-appropriate emotional expressions, focusing on fear and anger. We address three key questions: (1) How accurately can Stable Diffusion generate these emotions? (2) How does the specificity of textual prompts influence accuracy? (3) Are there observable biases in the generated expressions? Using scenarios from the EmotionBench dataset, we designed prompts to evoke specific emotional responses and analyzed the images with GPT-4V, an advanced emotion recognition model. Our findings indicate that while Stable Diffusion can generate fear with reasonable accuracy, it struggles with anger, suggesting a potential bias influenced by training data. Additionally, the model faces challenges in depicting emotions in complex contexts, such as scenes involving a person inside a car or in dark settings. We found no consistent impact of prompt specificity on accuracy, indicating other factors may influence performance. These insights highlight the need for diverse and high-quality training data and improved evaluation frameworks. Future research should incorporate a broader range of emotions, involve human evaluators, and standardize prompt specificity to enhance the reliability and comprehensiveness of AI-generated emotional expressions. Our study underscores the importance of these improvements to develop emotionally aligned and context-aware AI systems, ultimately enhancing human-computer interactions.

1 Introduction

Generative AI models have the potential to significantly enhance virtual communication and entertainment by accurately depicting emotional expressions. Emotions are fundamental to human interaction, and replicating these in AI-generated content can significantly enhance user experience and engagement [1]. Research by Affectiva [2] emphasizes the importance of emotional intelligence in human-computer interaction, focusing on algorithms that interpret facial expressions and map them to emotional states, thereby making virtual assistants more conversational and responsive.

While significant progress has been made in developing algorithms that can interpret direct emotional stimuli, the challenge of accurately recognizing and generating emotional nuance from contextual information remains largely underexplored. Studies by Matsumoto and Hwang [3] have shown that understanding the context in which emotions are expressed is crucial for accurate emotional interpretation. This highlights the need for generative AI models to consider the broader situational and cultural contexts that shape emotional expressions [4, 5, 6].

Despite the recent advancements of these models, generative AI models like Stable Diffusion often struggle to convey the intended emotional and contextual nuances in their outputs. While text-based models like GPT-3 have shown proficiency in emotional alignment, image generation models lag behind. Previous studies, such as the one by Huang et al. [7], have introduced the concept of emotional alignment and evaluated it in large language models using datasets like EmotionBench [8]. These studies provide a foundation for understanding how AI can align with human emotions. However, the capability of image generation models like Stable Diffusion to capture emotional expressions from contexts remains underexplored. Studies comparing GPT-3, DALL-E, and Stable Diffusion highlight significant variability in emotional alignment, indicating a need for further investigation [7, 9].

This study seeks to help further fill this gap by systematically evaluating how accurately

Stable Diffusion can depict context-appropriate emotional expressions, specifically focusing on fear and anger. By narrowing our scope to these two distinct emotions, we can provide more concrete insights into the capabilities and limitations of Stable Diffusion. Fear and anger were chosen due to their prevalence and significant impact on human interactions. Above that, fear and anger have distinctive facial expressions and physiological responses [10]. Their clear and recognizable features allow us to evaluate the model's accuracy and effectiveness more precisely. Consequently, we address the following key research questions:

1. How accurately can Stable Diffusion generate the emotions of fear and anger?
2. Are there observable biases in the generated emotional expressions?
3. How does the specificity of textual prompts influence the accuracy of the generated emotions?

Our findings indicate that while Stable Diffusion can generate fear with reasonable accuracy, it struggles significantly with generating appropriate expressions of anger. We also found that the model has a tendency to misrepresent anger as fear, which can be seen as a bias. Additionally, the analysis of prompt specificity shows no consistent impact on accuracy, suggesting that other factors may influence the model's performance. These insights highlight the need for more diverse and high-quality training data and improved evaluation frameworks.

Finally, we discuss the key aspects covered in the paper: In Section 2, we review the existing literature on the challenges of generating realistic emotional expressions, the concept of emotional alignment, and the effectiveness of GPT-4V in measuring emotions in AI-generated content. Section 3 details the methods used to design prompts and evaluate the emotional accuracy of the generated images. In Section 4, we present the results of our analysis, highlighting the strengths and weaknesses of Stable Diffusion in generating contextually appropriate emotional expressions. Section 5 discusses the implications of our findings, the limitations of our study, and potential areas for improvement. Section 6 addresses the ethical aspects of our research and the reproducibility of our methods. Finally, Section 7 concludes the paper by summarizing our findings and suggesting directions for future research.

2 Related Work

This section reviews the existing literature on three key areas: the challenges in generating realistic emotional expressions, the concept of emotional alignment and its evaluation, and the effectiveness of GPT-4V [11] in measuring emotions in AI-generated content. By examining these aspects, we aim to provide a comprehensive background that underscores the foundation of this research.

2.1 Challenges in emotion generation

Emotions are multifaceted and can be expressed through various channels such as facial gestures and body language [12]. The diversity of emotional expressions goes beyond the mere number of possible facial expressions. Physiologists have identified that facial muscles can form over twenty thousand unique expressions [13], illustrating the intricate nature of human emotions. Representing this feature space also involves capturing the context in which these expressions are perceived [4]. This inherent complexity of emotions and their expressions poses a significant challenge in the field of generative AI.

Diffusion models, such as Stable Diffusion, employ a process known as diffusion, which involves gradually refining the image over multiple steps [14]. Achieving the desired level of complexity through this iterative process demands intricate architectures capable of managing the high dimensionality of image data and capturing subtle emotional expressions.

Furthermore, these models require extensive training data to learn the countless ways emotions can be expressed. The diversity and volume of this data are crucial for the models to generalize well across different scenarios and emotional contexts. However, compiling such large, well-curated datasets is a significant challenge. The scarcity of high-quality, emotionally labeled datasets makes it difficult to train models that can accurately capture and generate nuanced emotional expressions [15].

Additionally, biases in training data present another significant challenge. According to Zhang et al. [16], text-to-image diffusion models can inherit and even amplify biases present in their training datasets. As a result, generative AI models may produce outputs that reinforce these biases, leading to skewed or inappropriate emotional expressions.

Our work addresses these challenges by specifically evaluating Stable Diffusion's ability to generate emotional expressions in diverse and complex scenarios, providing insights into areas that require improvement. By focusing on the generation of two distinct emotions-fear and anger-we also aim to identify and address potential biases in the model.

2.2 Emotional Alignment

While generative AI models are proficient in generating images from textual prompts, their capacity to precisely depict the intended emotional and contextual nuances of those prompts remains uncertain [17, 18, 19]. This limitation is particularly evident in the context of generating emotionally aligned content, where the AI must accurately reflect the intended emotions.

The concept of emotional alignment in AI refers to the ability of AI systems to accurately reflect the intended emotions in their outputs. This involves not only recognizing and replicating facial expressions but also understanding and capturing the context that gives these expressions their meaning. A pioneering study in this area is presented by the authors of "Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench" [7]. They established the concept of emotional alignment and conducted a comprehensive evaluation of emotion appraisal on different large language models (LLMs). Their work included collecting a diverse dataset of 428 situations encompassing 8 distinct negative emotions and establishing a human baseline through a user study involving 1,266 annotators from various demographics. The labeled dataset they put together, EmotionBench, is a comprehensive dataset designed to evaluate the emotional responses of LLMs. The dataset includes over 400 scenarios, and includes human evaluation results from over 1,200 subjects [8].

In another related study, "The Alignment of AI Emotions: Human Ratings of the Emotions Expressed by GPT-3, DALL-E and Stable Diffusion," the authors compared the emotional alignment of these AI models. They found significant variability in the models' abilities to produce aligned emotional expressions. GPT-3, for instance, produced more aligned emotional expressions compared to image generators like DALL-E and Stable Diffusion, potentially due to its larger model size and the textual nature of its outputs. This study highlights the ongoing challenges and progress in achieving emotional alignment across different types of AI models [9].

Our study builds on this by specifically examining how well Stable Diffusion can generate contextually appropriate emotional expressions, and how prompt specificity influences this capability. By investigating the model's performance across different levels of prompt detail, we aim to uncover the nuances in how specificity affects emotional accuracy. This research into prompt specificity aims to create a new area of exploration within the field, encouraging further studies to optimize and standardize prompt design for improved emotional alignment in generative AI models.

2.3 Measuring Emotions

Accurately measuring emotions in AI-generated content is crucial for evaluating the effectiveness of generative models like Stable Diffusion. This study employs GPT-4V, an advanced vision-language model, to assess the emotional expressions generated by these models. Recent research highlights the efficacy of GPT-4V in this domain.

Studies demonstrate that GPT-4V can accurately identify emotions without extensive fine-tuning, leveraging its comprehensive understanding of visual and textual data. This zero-shot learning approach enables GPT-4V to interpret a wide range of emotional expressions effectively, which is particularly beneficial for applications in diverse contexts [20]. Additionally, pilot evaluations support its ability to generate nuanced emotional responses from both visual and textual inputs, underscoring the model's potential to bridge the gap between human emotional expression and AI interpretation [21].

Research indicates that generalized large models like GPT-4V provide significant improvements in the accuracy and reliability of emotion recognition tasks. These models enhance the field by leveraging extensive training data and sophisticated architectures that allow for better generalization across diverse scenarios and datasets [22]. Furthermore, evaluations suggest that GPT-4V is effective in performing psychological assessments through visual affective computing, demonstrating its capability to analyze emotional states from visual cues [23]. This capability is crucial for applications requiring detailed emotional analysis and understanding, further establishing GPT-4V as a versatile tool for emotion recognition.

By leveraging the proven efficacy of GPT-4V in emotion recognition, this study aims to achieve a comprehensive assessment of Stable Diffusion's capabilities in generating contextually appropriate emotional expressions. This approach not only aligns with the research goals but also addresses the specific challenges of emotion recognition in AI-generated content.

3 Methods

To address the research questions, we employ a multi-step approach. First, we design prompts using scenarios from the EmotionBench dataset to evoke specific emotional expressions. These prompts are used to generate images using Stable Diffusion. The generated images are then evaluated using GPT-4V, a generative AI model that has demonstrated efficacy in emotion recognition [21, 20]. By analyzing the classification results, we can identify the model's strengths and weaknesses in generating contextually appropriate emotional expressions of fear and anger. This comprehensive evaluation not only helps in understanding the current capabilities of Stable Diffusion but also highlights areas for improvement and further research.

3.1 Prompt Design

To evaluate the model's capabilities, we employ textual prompts derived from the EmotionBench [8] dataset, which contains situations with verified emotional labels. For each emotion, we generate three prompts across twelve different scenarios, resulting in a total of 36 prompts per emotion. These prompts are designed to evoke specific emotional expressions. The specificity of each prompt is quantified by word count, allowing us to assess how the level of detail influences the model's output. This approach enables us to identify any potential biases or strengths the AI model may exhibit in depicting certain emotions or contextual details.

We utilize a custom version of OpenAI's GPT-4 [24] to create three prompts per scenario with varying levels of specificity: vague, moderately detailed, and highly detailed. The guidelines that were given to the GPT can be found in Figure 1.

Specific labels were added at the end of each prompt to provide clear instructions to the image generation model regarding the desired visual elements. The labels that were added at the end of each prompt were: "Face, Body and Front View". This method ensures that the generated images align more closely with an output that we can evaluate, enhancing the overall effectiveness of the prompts. By specifying visual components, we aim to reduce ambiguity and improve the accuracy of the generated images.

Figures 2 and 3 illustrate examples of the generated prompts and their respective generated images.

Guidelines for GPT-4

- Generate prompts that describe a person in a specific situation subtly suggesting an emotion, aiming to evoke a clear emotional response from the AI model.
- Maintain scenario integrity across different levels of specificity to ensure a controlled comparison of how detail level impacts the AI's ability to depict nuanced emotional expressions.
- Avoid mentioning any physical attributes, such as facial expressions, and focus on situational and environmental cues..
- Use straightforward language, avoiding complex terms that might not be well-represented in the model's training data.

Figure 1: GPT-4 Guidelines

3.2 Image Evaluation

The evaluation process involves using GPT-4V to assess the generated images. This method will analyze how accurately the images generated by Stable Diffusion reflect the intended emotions based on the given prompts. The classification results will help us validate the model's performance and highlight areas where the AI excels or needs improvement in emotional representation. To evaluate the emotional accuracy of the images generated by Stable Diffusion, the images are categorized into the following six classes:

- **Anger:** Images that depict a clear expression of anger.

Example Prompts Generated From EmotionBench Scenario
Emotion: "Anger"
EmotionBench Scenario: "You get home from the drive-thru and realize that you were given the wrong food."
Vague Prompt: "A person getting home and realizing their drive-thru order is wrong."
Moderate Detail Prompt: "A person arriving home from the drive-thru and discovering that their order is completely wrong."
High Detail Prompt: "A person excited to eat their meal after getting home from the drive-thru, opening the bag only to find out that the restaurant gave them someone else's order."

Figure 2: Example Prompts

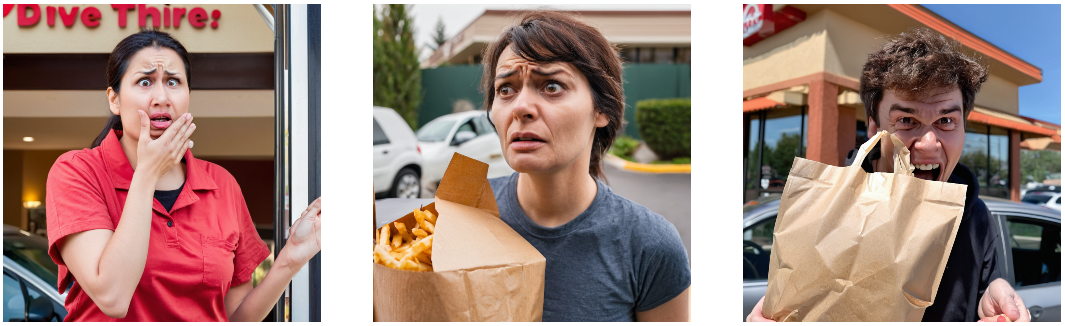


Figure 3: Example images generated from the respective vague, moderately detailed, and highly detailed prompts.

- **Fear:** Images that depict a clear expression of fear.
- **Other Negative Emotion:** Images that depict negative emotions other than anger or fear (e.g., sadness, disgust).
- **Other Positive Emotion:** Images that depict positive emotions (e.g., happiness, surprise).
- **No Emotion:** Images that do not depict any discernible emotion, possibly because the expression is ambiguous.
- **No person:** Images that do not contain a person to classify.

These classes were chosen to capture a comprehensive range of potential outputs from the model. By including categories for other negative and positive emotions, we aim to understand whether the model misclassifies the intended emotions of fear and anger into other emotional categories. The "No Emotion" and "No Person" categories help identify instances where the model fails to generate discernible emotional content or fails to generate a person at all. This categorization allows us to test the hypothesis that Stable Diffusion can accurately generate context-appropriate emotional expressions and helps us identify specific areas where the model may be prone to errors or biases.

To ensure accurate and consistent classification, GPT-4V is provided with a specific prompt alongside each image. The prompt used for GPT-4V is as follows:

> "System prompt: You are an expert in image analysis and emotional classification.
>
> User prompt: Given the following image, classify the depicted emotion of the central person into one of the following categories: anger, fear, other negative emotion, other positive emotion, no emotion, or no person. Include solely the classification in your response, you can only choose one category."

4 Results

The following section presents the findings from our study. We begin by detailing the experimental setup used to generate the images. Subsequently, we evaluate the model's performance in generating context-appropriate emotional expressions, specifically focusing on the emotions of fear and anger. We also examine the influence of prompt specificity on the accuracy of the generated emotions and explore any observable biases in the model's outputs.

4.1 Experimental Setup

For this study, we utilized Juggernaut XL, a version of the Stable Diffusion model SDXL. This model was chosen due to its advanced capabilities in generating high-resolution, photorealistic images, which are crucial for accurately depicting subtle emotional expressions. The results were generated using the webUI automatic1111 [25] through the app Rundiffusion [26].

Rundiffusion is an application that allows users to run diffusion models remotely in the cloud through the automatic1111 webUI. This setup provides several advantages, including the ability to leverage powerful cloud-based resources for image generation, which ensures high performance and scalability. Using Rundiffusion and automatic1111, we were able to efficiently generate images with consistent quality and detail. Images were produced using the default settings of the app to ensure consistency and reproducibility in the generation process.

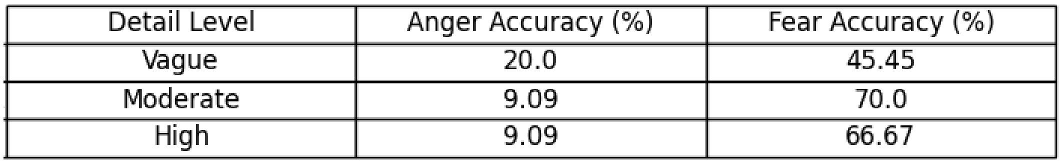
4.2 Results

Leveraging the methods outlined in Section 3, we address the following key Research Questions (RQs) to evaluate our study:

- **RQ1:** How accurately does Stable Diffusion generate context-appropriate emotional expressions, specifically focusing on fear and anger?
- **RQ2:** How does the specificity of textual prompts influence the accuracy of the generated emotions?
- **RQ3:** Are there observable biases in the generated emotional expressions?

RQ1 and RQ2: Accuracy of Emotion Generation and Influence of Prompt Specificity

Figure 4 summarizes the accuracy of Stable Diffusion in generating context-appropriate emotional expressions. The results are categorized based on the intended emotion (anger and fear) and the specificity of the prompts. The number of samples used for each accuracy computation is 36 per emotion. This amount is divided across each specificity level, resulting in 12 samples for each level within each emotion. It is important to note that for the accuracy of intended emotion, we excluded classifications where no emotion was displayed at all. This choice ensures a more accurate assessment of the model's performance in depicting the intended emotions, presenting the percentage of correct emotions against total emotions, rather than including ambiguous outputs.



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

Figure 4: Emotion accuracy and impact of prompt specificity.

The results indicate that Stable Diffusion struggles significantly with accurately depicting the emotion of anger. The accuracies for anger range from 9.09% to 20.0% across different levels of specificity. This suggests that the model has considerable difficulty in interpreting and generating expressions of anger from the given prompts.

In contrast, the model performs better in generating the emotion of fear. The accuracies for fear range from 45.45% to 70.0%, showing a more consistent ability to depict fear in response to the contextual information provided.

The highest recorded accuracy is 70% for fear with moderate detail prompts. However, the specificity of the anger prompts do not appear to follow the same trend. As shown in the table, while fear accuracy is highest with moderate detail, vague prompts perform the best for anger (20.0% for vague, and 9.09% for both moderate and high detail).

Overall, Stable Diffusion shows a relative strength in depicting fear compared to anger, regardless of prompt specificity. Further investigation is needed to understand the underlying factors affecting these results.

RQ3: Biases in Generated Emotions

Figures 5 and 6 show the distribution of generated emotions across different categories. The number of samples used for this analysis is consistent with the previous sections, with 36 prompts per emotion.

Figure 5 illustrates the distribution of classifications for scenarios where the expected emotion was anger. It is evident that the model predominantly generated fear-like expressions in these scenarios, indicating a significant bias towards fear even when anger was intended. Similarly, Figure 6 shows that fear is the most frequently generated emotion in scenarios designed to evoke fear. This suggests a tendency of the model to generate fear expressions in the contexts tested.

5 Discussion

Designing AI systems that can accurately generate emotional expressions is essential for enhancing human-computer interactions in various applications [27]. This study evaluated the capability of Stable Diffusion to depict context-appropriate emotions, specifically fear

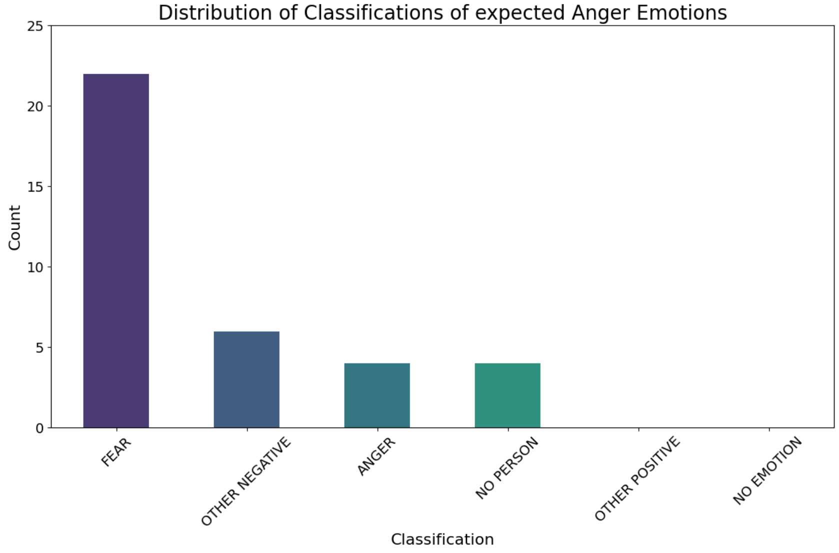


Figure 5: Distribution of classifications for expected angry emotions.

and anger, using scenarios from the EmotionBench dataset. By analyzing the generated images with GPT-4V, we aimed to uncover the strengths and limitations of the model.

5.1 Model Performance and Prompt Specificity

The results reveal that Stable Diffusion struggles with generating accurate expressions of anger, with the accuracy ranging between only 9.09% and 20.0%. In contrast, the model performed better in generating fear, with a measured accuracy between 45.45% and 70.0%. This disparity suggests that the model's ability to depict emotions may be influenced by imbalances in the training data, with a bias towards generating fear-like expressions, indicating that some emotions are underrepresented or depicted less accurately.

We also observed classifications of "no emotion" and "no person." The "no person" classification frequently occurred in contexts where the person was depicted inside a car. "No emotion" classifications were more prevalent in fear scenarios, often due to dark settings where facial features were not visible. This highlights the model's difficulty in handling complex contexts where visual elements are challenging to capture accurately.

The analysis shows that the highest accuracy for fear (70%) was achieved with moderate detail prompts. This suggests that providing a moderate amount of context could be optimal for the model. Vague prompts may lack sufficient information, while overly detailed prompts might overwhelm the model, indicating a need for a prompt with balanced details.

However, the specificity of the prompts does not appear to have a consistent impact on the accuracy of the generated emotions overall. While fear accuracy is highest with moderate detail, anger accuracy does not follow the same pattern. The accuracy for anger remains low across all levels of specificity, with the highest accuracy for vague prompts. This indicates variability in results that does not support a clear conclusion about the impact of prompt specificity.

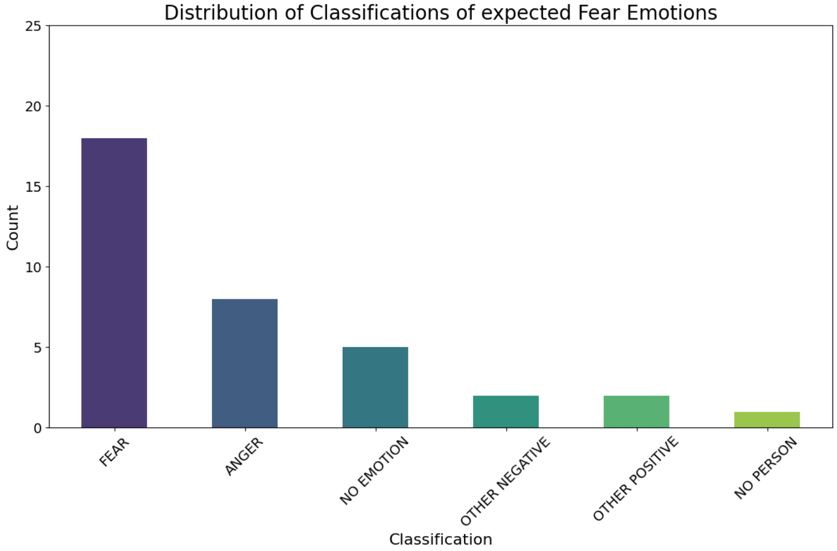


Figure 6: Distribution of classifications for expected fear emotions.

5.2 Implications and Recommendations

The findings from this study highlight several important considerations for improving the emotional evaluation of AI systems like Stable Diffusion. One notable limitation is the reliance on GPT-4V for emotion classification, which presents a narrow perspective. Human evaluators could provide a richer, more nuanced assessment of the model's performance, potentially uncovering subtleties that the AI might miss. Recent research supports this view, demonstrating that human judgment is essential for providing context and understanding complexities that AI models may not fully grasp [28].

Furthermore, the manipulation of prompt specificity using GPT-4 lacks a standardized definition of detail, which might affect the consistency of our results. These factors underscore the need for a more robust evaluation framework that includes human feedback.

Additionally, the limited range of emotions explored–only fear and anger–constrains our understanding of the model's capabilities. Future studies should include a broader spectrum of emotions to provide a more comprehensive evaluation.

Lastly, our sample size, based on 12 scenarios per emotion with 3 different prompts per scenario, while providing multiple data points, might introduce sampling bias, potentially skewing the overall findings. Increasing the number of unique scenarios tested could mitigate this issue.

In summary, the significant difference in accuracy between anger and fear, along with the model's bias towards generating fear-like expressions and its poor performance in recognizing and generating anger, underscores specific challenges within the model's performance. These limitations should be further explored by incorporating a wider range of emotions and involving human raters in the evaluation process, ensuring a more comprehensive assessment of the model. Additionally, standardizing the definition of prompt specificity might enhance the consistency and reliability of the results.

6 Responsible Research

This section outlines our approach to addressing potential biases, assessing potential harms and benefits, maintaining transparency and reproducibility, and considering cultural sensitivity. These elements are crucial to ensuring that our findings contribute positively and ethically to the development of AI systems capable of generating context-appropriate emotional expressions.

Although we did not identify a specific bias towards fear in the EmotionBench prompts, it is crucial to recognize that biases can exist in various forms. Potential biases within the input could influence the outcomes of AI models [29], particularly in how emotional expressions are depicted. We acknowledge the importance of continually assessing and addressing any biases to ensure fair and equitable performance of AI systems. Future research should stay alert in identifying and mitigating biases as they emerge.

The AI models used in this study have significant potential for both positive applications and misuse [30]. If not properly regulated, these systems could be used to manipulate emotions, spread misinformation, or create harmful content. For instance, AI-generated emotional expressions could be employed in deepfake videos to evoke false emotional responses, manipulate public opinion, or deceive individuals by creating realistic but fake interactions. This could lead to significant ethical and social issues, particularly in sensitive areas such as mental health and virtual communication. It is crucial to be aware about these potential risks.

To ensure transparency and facilitate reproducibility, we have provided comprehensive descriptions of our methods, including prompt design, experimental setup and evaluation criteria. Above that, the dataset that we used is publicly available, and our prompt set can be found in the appendix A, supporting the replication of our findings by other researchers.

While cultural sensitivity was not a primary focus of our testing framework, we are aware that emotions and their expressions can vary widely across different cultures [5]. Our study recognizes the well-known importance of cultural sensitivity [31] and strives to consider these differences in future research. Ensuring that AI models are inclusive and culturally aware remains a critical goal for developing universally applicable systems.

7 Conclusions and Future Work

7.1 Conclusions

This study evaluates Stable Diffusion's ability to generate context-accurate emotional expressions, focusing on fear and anger. Our findings indicate that while Stable Diffusion can generate fear with reasonable accuracy, it struggles significantly with generating appropriate expressions of anger in the given contexts. This discrepancy suggests a potential bias in the model, possibly due to imbalances in the training data where fear-related expressions are better represented.

Furthermore, the analysis of prompt specificity shows no consistent impact on accuracy, suggesting that other factors may influence the model's performance. This variability indicates that a one-size-fits-all approach to prompt specificity might not be effective.

Additionally, the model faces challenges in accurately depicting emotions in complex contexts, such as scenes involving a person inside a car or in dark settings. These limitations highlight that Stable Diffusion's current capabilities are constrained by the quality and diversity of its training data and the complexity of the scenarios it needs to depict.

7.2 Future Work

Building on the insights from this study, future research should aim to expand the evaluation framework and address the identified limitations. A crucial next step involves incorporating a broader range of emotions beyond fear and anger to provide a more comprehensive understanding of the model's capabilities. This would help determine whether the observed biases and challenges are consistent across other emotional expressions or specific to fear and anger.

Moreover, involving human raters in the evaluation process could offer deeper insights into the model's performance. Human evaluators can capture subtleties and nuances in emotional expressions that automated systems might miss, enriching the evaluation framework and making it more reflective of real-world perceptions of emotion.

Another important area for future research is the standardization of prompt specificity. Developing a clear and consistent method for defining prompt specificity will ensure more reliable and reproducible results. By standardizing the level of detail provided in prompts, researchers can better understand how prompt specificity affects the model's ability to generate accurate emotional expressions and identify any systematic biases that may arise from different prompt styles.

Furthermore, future work should consider testing other advanced models to compare their performance in generating context-appropriate emotional expressions. Models such as DALL-E 2 [32], Imagen by Google [33], and MidJourney [34] offer different architectures and training datasets, which may provide valuable insights into the strengths and limitations of various approaches to emotional AI.

Exploring these directions will provide a more nuanced and comprehensive assessment of generative AI models like Stable Diffusion. By addressing these aspects, future studies can enhance the evaluation framework, ultimately contributing to the development of more emotionally aligned and context-aware AI systems. These advancements are vital for improving human-computer interactions and ensuring that AI systems can effectively support applications requiring nuanced emotional understanding and expression.

Acknowledgements

This research benefited from the assistance of the OpenAI GPT-4 language model, which provided support in refining the text and grammar. The authors critically reviewed and edited all content generated by GPT-4 to ensure accuracy and coherence.

References

[1] R. W. Picard, *Affective computing.* MIT press, 2000.

[2] Affectiva, "Emotion ai," https://www.affectiva.com, accessed: May 22, 2024.

[3] D. Matsumoto and H. Sung Hwang, "Judging faces in context," *Social and Personality Psychology Compass*, vol. 4, no. 6, pp. 393–402, 2010.

[4] K. H. Greenaway, E. K. Kalokerinos, and L. A. Williams, "Context is everything (in emotion research)," *Social and Personality Psychology Compass*, vol. 12, no. 6, p. e12393, 2018.

[5] D. Matsumoto, “Culture and emotion,” *The handbook of culture and psychology*, pp. 171–194, 2001.

[6] M. Kayyal, S. Widen, and J. Russell, “Context Is More Powerful Than We Think: Contextual Cues Override Facial Cues Even for Valence,” *Emotion (Washington, D.C.)*, vol. 15, Feb. 2015.

[7] 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, arXiv:2308.03656 [cs]. [Online]. Available: http://arxiv.org/abs/2308.03656

[8] CUHK-ARISE, “Emotionbench/situations at main,” https://github.com/CUHK-ARISE/EmotionBench/tree/main/situations, accessed: May 22, 2024.

[9] 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.

[10] P. Ekman, “An argument for basic emotions,” *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[11] OpenAI, “Gpt-4v: Combining language and vision,” 2024, accessed: 2024-06-03. [Online]. Available: https://openai.com/research/gpt-4

[12] B. De Gelder, A. W. de Borst, and R. Watson, “The perception of emotion in body expressions,” *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 2, pp. 149–158, 2015.

[13] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, 1978.

[14] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, “Diffusion models: A comprehensive survey of methods and applications,” *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[15] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “From facial expression recognition to interpersonal relation prediction,” *International Journal of Computer Vision*, vol. 126, pp. 550–569, 2018.

[16] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, “Text-to-image diffusion models in generative ai: A survey. arxiv 2023,” *arXiv preprint arXiv:2303.07909.*

[17] Y. Wang, S. Shen, and B. Y. Lim, “Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3581402

[18] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie, “Large language models understand and can be enhanced by emotional stimuli,” *arXiv preprint arXiv:2307.11760*, 2023.

[19] C. Li, J. Wang, Y. Zhang, K. Zhu, X. Wang, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie, “The good, the bad, and why: Unveiling emotions in generative ai,” 2023.

[20] Z. Lian, L. Sun, H. Sun, K. Chen, Z. Wen, H. Gu, B. Liu, and J. Tao, “Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition,” *Information Fusion*, vol. 108, p. 102367, 2024.

[21] Z. Elyoseph, E. Refoua, K. Asraf, M. Lvovsky, Y. Shimoni, and D. Hadar-Shoval, “Capacity of generative ai to interpret human emotions from visual and textual data: Pilot evaluation study,” *JMIR Ment Health*, vol. 11, p. e54369, Feb 2024. [Online]. Available: https://mental.jmir.org/2024/1/e54369

[22] Z. Zhang, L. Peng, T. Pang, J. Han, H. Zhao, and B. W. Schuller, “Refashioning emotion recognition modelling: The advent of generalised large models,” *IEEE Transactions on Computational Social Systems*, 2024.

[23] H. Lu, X. Niu, J. Wang, Y. Wang, Q. Hu, J. Tang, Y. Zhang, K. Yuan, B. Huang, Z. Yu *et al.*, “Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing,” *arXiv preprint arXiv:2403.05916*, 2024.

[24] OpenAI, “Gpt-4: Openai’s language model,” OpenAI Website, 2024, accessed: 2024-05-01. [Online]. Available: https://www.openai.com/models/gpt-4

[25] automatic1111, “stable-diffusion-webui,” https://github.com/AUTOMATIC1111/stable-diffusion-webui, 2023, accessed: 2024-06-10.

[26] Rundiffusion, “Rundiffusion,” https://app.rundiffusion.com, 2023, accessed: 2024-06-10.

[27] P. Priya, M. Firdaus, G. V. Singh, and A. Ekbal, “Affective computing for social good applications: Current advances, gaps and opportunities in conversational setting,” in *European Conference on Information Retrieval.* Springer, 2024, pp. 375–380.

[28] Y. Liu, H. Zhou, Z. Guo, E. Shareghi, I. VuliÄ, A. Korhonen, and N. Collier, “Aligning with human judgement: The role of pairwise preference in large language model evaluators,” 2024.

[29] A. Liusie, P. Manakul, and M. J. F. Gales, “Mitigating word bias in zero-shot prompt-based classifiers,” 2023.

[30] L. PÃ¶hler, V. Schrader, A. Ladwein, and F. von Keller, “A technological perspective on misuse of available ai,” 2024.

[31] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[32] OpenAI, “Dall-e 2: Creating images from text,” https://www.openai.com/dall-e-2, 2022, accessed: 2024-06-10.

[33] G. Research, “Imagen: Text-to-image diffusion models,” https://imagen.research.google, 2022, accessed: 2024-06-10.

[34] MidJourney, “Midjourney,” https://www.midjourney.com, 2022, accessed: 2024-06-10.

A Textual Prompts Used in the Study

A.1 Prompts for Anger

Scenario 1: A person nearly gets hit by someone backing their car out of the driveway.

- **Vague Prompt:** A person that nearly gets hit by someone backing their car out of the driveway. Face, body, front view.
- **Moderate Detail Prompt:** A person driving through a residential area when someone suddenly backs their car out of a driveway and nearly collides with their vehicle. Face, body, front view.
- **High Detail Prompt:** A person driving through a quiet residential area when a car suddenly backs out of a driveway without warning, forcing them to swerve to avoid a collision. Face, body, front view.

Scenario 2: A person seeing their new roommate leave the kitchen messy after a gathering.

- **Vague Prompt:** A person seeing their new roommate leave the kitchen messy after a gathering. Face, body, front view.
- **Moderate Detail Prompt:** A person finding the kitchen in a mess with dirty dishes and leftover food after their roommate had friends over. Face, body, front view.
- **High Detail Prompt:** A person returning home after a long day to find the kitchen in chaos with dirty dishes and food scraps everywhere because their new roommate didn't clean up after hosting friends. Face, body, front view.

Scenario 3: A person being bumped into at the mall without an apology.

- **Vague Prompt:** A person being bumped into at the mall without an apology. Face, body, front view.
- **Moderate Detail Prompt:** A person walking through the crowded mall when someone rudely bumps into them and continues on without apologizing. Face, body, front view.
- **High Detail Prompt:** A person navigating through a busy mall when someone roughly bumps into them, causing them to almost drop their bags, and walks away without a word of apology. Face, body, front view.

Scenario 4: A person getting home and realizing their drive-thru order is wrong.

- **Vague Prompt:** A person getting home and realizing their drive-thru order is wrong. Face, body, front view.
- **Moderate Detail Prompt:** A person arriving home from the drive-thru and discovering that their order is completely wrong. Face, body, front view.

- **High Detail Prompt:** A person excited to eat their meal after getting home from the drive-thru, opening the bag only to find out that the restaurant gave them someone else's order. Face, body, front view.

Scenario 5: A person stuck behind a slow driver on an open road.

- **Vague Prompt:** A person stuck behind a slow driver on an open road. Face, body, front view.
- **Moderate Detail Prompt:** A person on an otherwise open road, finding themselves stuck behind a driver going well below the speed limit. Face, body, front view.
- **High Detail Prompt:** A person driving on an open road, becoming increasingly frustrated as they get stuck behind a car moving at a snail's pace, with no chance to overtake. Face, body, front view.

Scenario 6: A person at a store waiting to be helped, while the clerks are talking and ignoring them.

- **Vague Prompt:** A person at a store waiting to be helped, while the clerks are talking and ignoring them. Face, body, front view.
- **Moderate Detail Prompt:** A person waiting for assistance at a store, while the clerks continue chatting amongst themselves and ignore their presence. Face, body, front view.
- **High Detail Prompt:** A person standing at the counter of a store, growing increasingly irritated as the clerks engage in their own conversation, blatantly ignoring their repeated attempts to get attention. Face, body, front view.

Scenario 7: A person whose spouse didn't do something they promised to take care of.

- **Vague Prompt:** A person whose spouse didn't do something they promised to take care of. Face, body, front view.
- **Moderate Detail Prompt:** A person whose significant other fails to complete a task they had promised to take care of. Face, body, front view.
- **High Detail Prompt:** A person finding out that their spouse didn't take care of an important task they promised to handle, leaving them to deal with the consequences at the last minute. Face, body, front view.

Scenario 8: A person being talked down to.

- **Vague Prompt:** A person being talked down to. Face, body, front view.
- **Moderate Detail Prompt:** A person being spoken to in a condescending manner, making them feel belittled. Face, body, front view.
- **High Detail Prompt:** A person during a conversation, being repeatedly talked down to, using a condescending tone and dismissive language that makes them feel inferior. Face, body, front view.

Scenario 9: A person whose family doesn't take their career seriously.

- **Vague Prompt:** A person whose family doesn't take their career seriously. Face, body, front view.
- **Moderate Detail Prompt:** A person whose family makes comments that show they don't take their education or career seriously. Face, body, front view.
- **High Detail Prompt:** A person at a family gathering, overhearing their relatives making dismissive remarks about their educational and career choices, showing their lack of support and understanding. Face, body, front view.

Scenario 10: A person in a group project, finding out that a coworker has not completed their responsibilities.

- **Vague Prompt:** A person in a group project, finding out that a coworker has not completed their responsibilities. Face, body, front view.
- **Moderate Detail Prompt:** A person discovering that a coworker in their group project has neglected to complete their assigned tasks. Face, body, front view.
- **High Detail Prompt:** A person during a group project review, realizing that one of their coworkers has completely ignored their responsibilities, jeopardizing the success of the entire project. Face, body, front view.

Scenario 11: A person being criticized by their friend in front of others.

- **Vague Prompt:** A person being criticized by their friend in front of others. Face, body, front view.
- **Moderate Detail Prompt:** A person whose friend makes critical remarks about them in front of a group of people. Face, body, front view.
- **High Detail Prompt:** A person in a social setting, being openly criticized by their friend in front of several other people, leaving them embarrassed and hurt. Face, body, front view.

Scenario 12: A person being criticized by their family for a decision they've made.

- **Vague Prompt:** A person being criticized by their family for a decision they've made. Face, body, front view.
- **Moderate Detail Prompt:** A person whose family members express their disapproval of a decision they've made. Face, body, front view.
- **High Detail Prompt:** A person during a family discussion, whose relatives harshly criticize a significant decision they made, questioning their judgment and making them feel unsupported. Face, body, front view.

A.2 Prompts for Fear

Scenario 1: A person walking alone when a stranger jumps out of the bushes to attack.

- **Vague Prompt:** A person walking alone when a stranger jumps out of the bushes to attack. Face, body, front view.
- **Moderate Detail Prompt:** A person walking alone in a familiar area when a stranger jumps out of the bushes to attack. Face, body, front view.
- **High Detail Prompt:** A person walking alone in an isolated but familiar area when a menacing stranger suddenly jumps out of the bushes to attack them. Face, body, front view.

Scenario 2: A person alone in an elevator at night when a stranger rushes in to attack.

- **Vague Prompt:** A person alone in an elevator at night when a stranger rushes in to attack. Face, body, front view.
- **Moderate Detail Prompt:** A person alone in an elevator late at night when a stranger rushes in to attack and block the door. Face, body, front view.
- **High Detail Prompt:** A person alone in an elevator late at night when the doors open and a menacing stranger rushes in to attack them, blocking the door. Face, body, front view.

Scenario 3: A person stopped at a traffic signal when a stranger bangs on their car window and yells.

- **Vague Prompt:** A person stopped at a traffic signal when a stranger bangs on their car window and yells. Face, body, front view.
- **Moderate Detail Prompt:** A person alone in their car stopped at a traffic signal when a stranger bangs on their window and yells threatening things. Face, body, front view.
- **High Detail Prompt:** A person alone in a car on their way home when an angry stranger begins banging on their car window and yelling threatening things at them at a traffic signal. Face, body, front view.

Scenario 4: A person driving on a two-lane road when a car tailgates and honks at them aggressively.

- **Vague Prompt:** A person driving on a two-lane road when a car tailgates and honks at them aggressively. Face, body, front view.
- **Moderate Detail Prompt:** A person on a two-lane road when a car dangerously tailgates them, honking aggressively as it follows too closely. Face, body, front view.
- **High Detail Prompt:** A person driving along a two-lane road when they see in their rearview mirror that a car is dangerously tailgating them, honking their horn aggressively while continuing to follow too closely. Face, body, front view.

Scenario 5: A person walking through an unfamiliar area past midnight when a man shoves them.

- **Vague Prompt:** A person walking through an unfamiliar area past midnight when a man shoves them. Face, body, front view.
- **Moderate Detail Prompt:** A person walking through an unfamiliar area past midnight when they accidentally run into a man who becomes angry and shoves them. Face, body, front view.
- **High Detail Prompt:** A person walking through an unfamiliar part of town past midnight when they accidentally run into a man who becomes angry and shoves them. Face, body, front view.

Scenario 6: A person in an empty parking lot when an acquaintance begins to shove and push them.

- **Vague Prompt:** A person in an empty parking lot when an acquaintance begins to shove and push them. Face, body, front view.
- **Moderate Detail Prompt:** A person talking in an empty parking lot when an acquaintance starts shoving and pushing them, leaving them unsure if the acquaintance is serious or kidding. Face, body, front view.
- **High Detail Prompt:** A person standing around and talking in an empty parking lot with someone they do not know well when the acquaintance begins to shove and push them, leaving them unsure if they are serious or just kidding around. Face, body, front view.

Scenario 7: A person in a park at night when they see a stranger with a knife approaching them.

- **Vague Prompt:** A person in a park at night when they see a stranger with a knife approaching them. Face, body, front view.
- **Moderate Detail Prompt:** A person at night in a park when they see a menacing stranger with a knife about 30 ft away approaching them. Face, body, front view.
- **High Detail Prompt:** A person outside in a park area at night when they see a menacing stranger with a knife about 30 ft away directly approaching them, obviously planning to attack. Face, body, front view.

Scenario 8: A person exiting an empty campus building at night when they feel a hand grab their arm.

- **Vague Prompt:** A person exiting an empty campus building at night when they feel a hand grab their arm. Face, body, front view.
- **Moderate Detail Prompt:** A person late at night exiting an empty campus building when they suddenly feel a hand grab their arm. Face, body, front view.
- **High Detail Prompt:** A person alone exiting an empty campus building late at night when they feel a hand grab their arm. Face, body, front view.

Scenario 9: A person sleeping at night when they wake up to a suspicious noise in the dark.

- **Vague Prompt:** A person sleeping at night when they wake up to a suspicious noise in the dark. Face, body, front view.
- **Moderate Detail Prompt:** A person alone in bed at night when they wake up to a suspicious noise in the dark. Face, body, front view.
- **High Detail Prompt:** A person sleeping in bed during the night when they suddenly wake up thinking they heard a suspicious noise. It is dark and they are alone. Face, body, front view.

Scenario 10: A person alone at home when they get a call from someone saying they are outside their house.

- **Vague Prompt:** A person alone at home when they get a call from someone saying they are outside their house. Face, body, front view.
- **Moderate Detail Prompt:** A person alone at home at night when they get a call from an unfamiliar voice saying they are outside their house. Face, body, front view.
- **High Detail Prompt:** A person alone at home one night about to go to bed when the phone rings. They answer it, and an unfamiliar voice on the other end tells them they are right outside their house and hangs up. Face, body, front view.

Scenario 11: A person coming home when they find a package by the mailbox that ticks when they open it.

- **Vague Prompt:** A person coming home when they find a package by the mailbox that ticks when they open it. Face, body, front view.
- **Moderate Detail Prompt:** A person finding an unexpected package by the mailbox when they notice a faint ticking sound as they open it. Face, body, front view.
- **High Detail Prompt:** A person coming home one day when they find an unexpected shoebox-sized package waiting for them by the mailbox. As they sit down to open it, they notice a faint ticking sound coming from inside the package. Face, body, front view.

Scenario 12: A person alone at home when they hear movement and whispering outside their window.

- **Vague Prompt:** A person alone at home when they hear movement and whispering outside their window. Face, body, front view.
- **Moderate Detail Prompt:** A person reading a book alone at night when they hear movement and whispering outside their window. Face, body, front view.
- **High Detail Prompt:** A person alone at home one night when they have settled down to read a book and hear some movement right outside their window. They cannot see anything, but when they listen more closely, it sounds like people whispering. Face, body, front view.