

Enhancing Stress Detection on Social Media Through Multi-Modal Fusion of Text and Synthesized Visuals

Efstathia Soufleri

Archimedes, Athena Research Center
Greece
e.soufleri@athenarc.gr

Sophia Ananiadou

The University of Manchester
Manchester, UK
Archimedes, Athena Research Center
Greece
sophia.ananiadou@manchester.ac.uk

Abstract

Social media platforms generate an enormous volume of multi-modal data, yet stress detection research has predominantly relied on text-based analysis. In this work, we propose a novel framework that integrates textual content with synthesized visual cues to enhance stress detection. Using the generative model DALL-E, we synthesize images from social media posts, which are then fused with text through the multi-modal capabilities of a pre-trained CLIP model, which encodes both text and image data into a shared semantic space. Our approach is evaluated on the Dreddit dataset, where a classifier trained on frozen CLIP features achieves 94.90% accuracy, and full fine-tuning further improves performance to 98.41%. These results underscore the integration of synthesized visuals with textual data not only enhances stress detection but also offers a robust method over traditional text-only methods, paving the way for innovative approaches in mental health monitoring and social media analytics.

1 Introduction

Social media has emerged as a pervasive platform for personal expression, generating enormous volumes of data that encompass both textual and visual modalities (Baltrušaitis et al., 2018; Mouzannar et al., 2018; Abousaleh et al., 2020). This rich, heterogeneous data offers unprecedented opportunities for understanding human behavior and mental health. However, prevailing stress detection research has largely focused on text-based analysis, overlooking the potential for complementary affective cues that can be inferred or synthesized into visual representations.

Recent advances in multi-modal machine learning have shown that integrating diverse data sources can significantly enhance performance on affective and behavioral tasks (Song et al., 2024; Ieracitano et al., 2020; Amal et al., 2022; Zhang et al., 2020). At the same time, generative models such

as DALL-E have opened new avenues for synthesizing high-quality visuals from textual descriptions (Ramesh et al., 2021; Khachatryan et al., 2023; Tewel et al., 2022). This proliferation of data prompts an essential question: How can the fusion of synthesized visual data with traditional text data improve the accuracy and effectiveness of stress detection algorithms?

In this work, we introduce a novel multi-modal framework that leverages both text and synthesized images for stress detection. Specifically, we generate images from social media posts using DALL-E (Ramesh et al., 2021) and integrate these visuals with text via the robust joint embedding space provided by a pre-trained CLIP model (Radford et al., 2021). We evaluate our approach on the Dreddit dataset (Turcan and McKeown, 2019), a collection of social media posts annotated to indicate whether the person who wrote the post suffers from stress or not. Our experiments demonstrate that a classifier trained on frozen CLIP features achieves 94.90% accuracy, while full fine-tuning further elevates performance to 98.41%. These results indicate that synthesized visuals capture subtle emotional and contextual cues that are absent from text alone, thereby significantly enhancing detection accuracy.

Our contributions are threefold:

1. We propose a novel multi-modal framework that fuses text and synthesized visuals to address the limitations of traditional text-only stress detection methods.
2. We demonstrate the effectiveness of our approach on the Dreddit dataset (Turcan and McKeown, 2019), achieving state-of-the-art performance through both classifier-only training and full fine-tuning strategies.
3. We provide an in-depth analysis of the impact of multi-modal fusion on capturing nuanced affective signals, laying the groundwork for

future research in mental health monitoring using social media data.

2 Related Work

Stress detection on social media has traditionally been approached using text-based methods. Early studies primarily relied on lexicon-based techniques and classical machine learning algorithms to identify linguistic markers of stress and anxiety in user-generated content (De Choudhury et al., 2013; Aldarwish and Ahmad, 2017; Biswas and Hasiija, 2022). More recent approaches have employed deep learning architectures, such as recurrent neural networks (Salehinejad et al., 2017) and transformer-based architectures (Vaswani et al., 2017; Ji et al., 2022; Yang et al., 2024; Shi et al., 2024), to capture complex syntactic and semantic patterns from text. Despite these advancements, text-only methods may fail to capture affective or contextual information that can be made more salient through synthesized visual representations.

The growing interest in multi-modal learning has spurred research into integrating multiple data sources to improve performance on affective and behavioral tasks. Several studies have demonstrated that fusing textual and visual information can enhance emotion recognition (Kosti et al., 2017; Cowie et al., 2001; Zhu et al., 2025) and sentiment analysis (Baltrušaitis et al., 2018; Wankhade et al., 2022). For instance, multi-modal architectures that combine convolutional neural networks (Li et al., 2021) for image analysis with language models for text have shown improved accuracy over single-modality approaches (Mittal et al., 2018; You et al., 2015; Feng et al., 2025; Devlin et al., 2018; Liu et al., 2019). However, the application of multi-modal techniques to stress detection remains relatively underexplored.

Generative models have further broadened the horizons of multi-modal research. Models such as DALL-E have shown impressive capabilities in synthesizing high-quality images from textual prompts (Ramesh et al., 2021; Zhou et al., 2023), thereby providing a novel means to augment datasets that lack explicit visual content. Concurrently, models like CLIP have established robust joint embedding spaces that effectively capture the semantic relationships between images and text (Radford et al., 2021; Qiao et al., 2019; Wang et al., 2023; Zhong et al., 2021; Gu et al., 2023; Wang et al., 2021). These innovations have paved the way for leverag-

ing synthesized visuals to complement textual data, offering new insights into affective states that may not be fully captured by text alone.

Prior work in mental health has shown that linguistic patterns in social media (e.g., first-person pronouns, hopelessness, negative tone) indicate stress, anxiety, or depression (De Choudhury et al., 2013; Cohan et al., 2018), and visual cues (e.g., expressions, colors, context) also reflect affective states (Abousaleh et al., 2020). Building on this, we hypothesize that even synthesized images—when guided by affect-sensitive prompts—can offer complementary signals for stress detection.

Our work builds on these lines of research by integrating synthesized visuals with text-based analysis for stress detection. Our work employs generative image synthesis in conjunction with a multi-modal representation framework for this task. By fusing the complementary strengths of DALL-E and CLIP, we aim to address the limitations of traditional text-only approaches and provide a more holistic understanding of stress as expressed on social media.

3 Methodology

In this section, we describe our multi-modal framework for stress detection, which integrates synthesized visual cues with textual information. Our approach consists of two stages: image generation, and multi-modal representation with CLIP.

3.1 Image Generation

To enrich textual data, we use the generative capabilities of DALL-E 3, an advanced version of the DALL-E model (Ramesh et al., 2021). This model synthesizes images closely aligned with textual descriptions. The process begins with the input of a text prompt into a specialized *text encoder*. This text encoder is adept at converting the textual information into a high-dimensional representation space (*text encoding*), aiming to capture the core semantic content of the prompt (Figure 1).

Following this, a component known as the *diffusion prior* takes over, which is a crucial part of the model’s architecture. The prior is responsible for mapping the text-encoded semantic representation to a corresponding *image encoding*. This image encoding is designed to retain the semantic content conveyed by the text, ensuring that the generated images reflect the intended themes and elements of the input prompt.

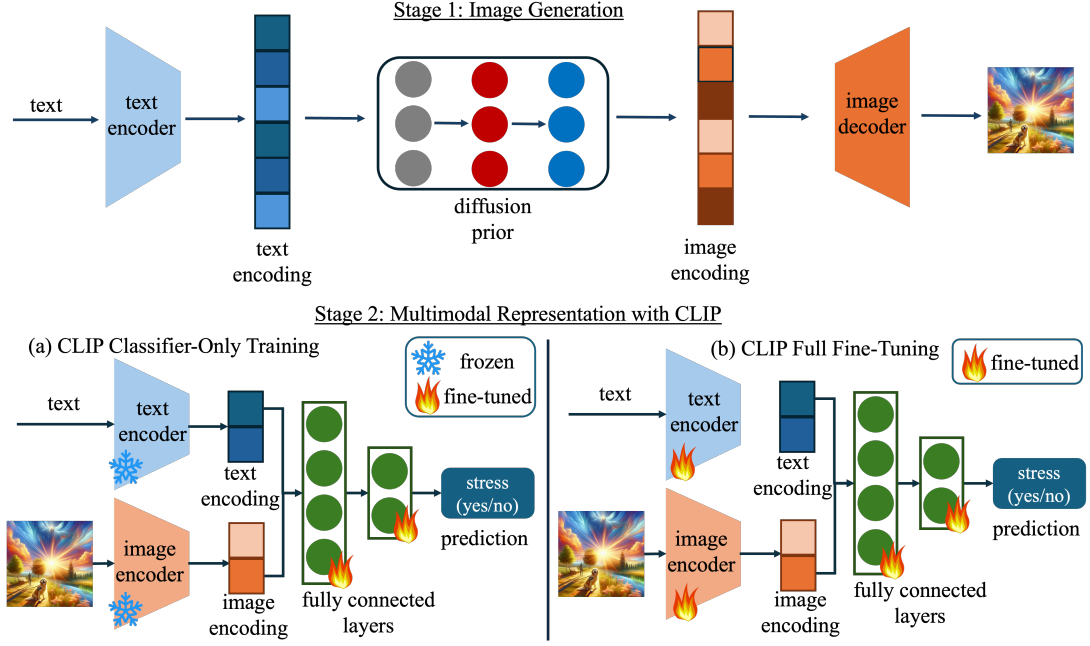


Figure 1: Methodology overview. (1) image generation: the posts (text) and the corresponding prompt are converted into images that visually represent the text’s semantic content, (2) multi-modal representation with CLIP: the images, alongside the original text, are processed through CLIP to form a joint embedding space, used for stress detection. The CLIP Classifier-Only Training strategy fine-tunes the classifier (fully connected layers) while keeping the CLIP base model (text and image encoder) frozen. The CLIP Full Fine-Tuning strategy fine-tunes both the classifier and the CLIP base model. This process leverages both textual and visual data to enhance detection accuracy.

The final step in the image generation process involves an *image decoder*. This decoder uses the image encoding to stochastically generate the final visual output. The resulting *image* is a visual representation of the semantic information encoded from the initial text prompt, materializing as a synthetic image that complements the textual data in our multi-modal stress detection framework. By leveraging this advanced image synthesis process, we ensure that the generated visuals are both semantically relevant and visually coherent, providing a robust foundation for further multi-modal analysis.

3.2 Multi-modal Representation with CLIP

We employ the pre-trained CLIP model (Radford et al., 2021) to facilitate a robust multi-modal representation, leveraging its capacity to encode both text and images into a shared joint embedding space. Each text sample is processed by the *text encoder* to extract textual features, while corresponding synthesized images are preprocessed and passed through the *image encoder*. The features from both modalities are normalized and concatenated to form a joint representation (Figure 1). This embedding captures complementary affective cues from both textual and visual data, enhancing our

ability to detect stress signals on social media.

To effectively train our model, we adopt two training strategies:

1. **CLIP Classifier-Only Training:** In this approach, we keep the pre-trained CLIP base frozen, focusing training efforts solely on the attached *classifier*. This method benefits from the robustness of the existing multi-modal embeddings, avoiding alterations to the underlying representations and ensuring stability.
2. **CLIP Full Fine-Tuning:** Alternatively, we engage in full fine-tuning of both the CLIP model and the *classifier*. This strategy allows the entire network to adapt more comprehensively to the domain-specific nuances of stress-related content, improving detection accuracy by refining the joint embedding space to better capture subtle emotional nuances.

This integrated methodology not only leverages generative image synthesis to augment textual information but also strategically fuses these modalities in a joint embedding space. The approach is designed to enhance the detection of nuanced affective signals that are pivotal for accurate stress detection on social media platforms.

Model	Accuracy (%)	Weighted F1 (%)
MentalRoBERTa	96.14	94.24
MentalBERT	69.32	78.75
RoBERTa-base	96.14	94.24
BERT-base	96.14	94.67
CLIP Classifier-Only Training (Ours)	94.90	92.42
CLIP Full Fine-Tuning (Ours)	98.41	98.27

Table 1: Performance comparison of our approach with general-purpose and mental health-specific models on the Dreddit dataset.

4 Experiments and Results

In this section, we outline our experimental setup, present results, and discuss findings.

4.1 Experimental Setup

The Dreddit dataset (Turcan and McKeown, 2019) comprises 2,837 training samples and 414 testing samples, where each sample is a social media post accompanied by a binary label indicating the presence or absence of stress. The posts are drawn from mental health-related subreddits such as r/depression, r/anxiety, and r/relationships. Each post was annotated through crowdsourced judgments, with three annotators per instance and majority voting used to determine the final label. The dataset is approximately balanced across the two classes. For each post, we generate a corresponding synthetic image using DALL-E 3. The hyperparameters reported in the Appendix. Our experiments compare the following models:

- **CLIP Classifier-Only Training (Ours):** Classifier-only training where the pre-trained CLIP model is kept frozen while only the classifier is trained.
- **CLIP Full Fine-Tuning (Ours):** Full fine-tuning of both the CLIP model and the classifier on the Dreddit dataset.
- **Text-Only Baselines:** Pre-trained discriminative language models which are either general purpose (RoBERTa-base, BERT-base (Devlin et al., 2018; Liu et al., 2019)) or finetuned for mental health applications (MentalRoBERTa, MentalBERT (Ji et al., 2022)).

4.2 Results

Table 1 reports the accuracy and weighted F1 scores for our proposed models and the text-only baselines. Our CLIP Classifier-Only Training model achieves an accuracy of 94.90% with a weighted F1 score of 92.42%, while the CLIP Full Fine-Tuning model reaches 98.41% accuracy and

Modality	Accuracy (%)	Weighted F1 (%)
Image-Only	95.22	93.17
Text-Only	96.82	96.31
Image + Text	98.41	98.27

Table 2: Ablation study of our method comparing image-only, text-only, and combined multi-modal model.

98.27% weighted F1 score. In comparison, the text-only models yield competitive performance for MentalRoBERTa, RoBERTa-base, and BERT-base (accuracy around 96.14% and weighted F1 around 94%), whereas MentalBERT underperforms. The results demonstrate that full fine-tuning of our multi-modal framework (CLIP Full Fine-Tuning) leads to a substantial improvement in performance over classifier-only training, highlighting the benefit of adapting the joint image-text representations to stress detection. Furthermore, our approach achieves competitive performance compared to strong text-only baselines, while offering the added advantage of leveraging synthesized visual cues. Even though our results demonstrate strong performance gains, we acknowledge that we have not conducted statistical significance testing across multiple random seeds. Future work will incorporate such evaluations to better assess the robustness of our findings.

4.3 Ablation Study: Modality Contributions

To better understand the contribution of each modality, we performed an ablation study by evaluating our model trained using only the synthesized images, only the textual data, and the fusion of both modalities Table 2. The image-only model, which relies solely on visual cues extracted from generated images, achieved an accuracy of 95.22% and a validation weighted F1 score of 93.17%. The text-only model, using only the original social media posts, reached an accuracy of 96.82% and a validation weighted F1 score of 96.31%. Notably, when both modalities are integrated, our multi-modal framework achieves significantly improved performance, with an accuracy of 98.41% and a validation weighted F1 score of 98.27%. These findings indicate that while the text-only model is already highly effective, the addition of synthesized visual information provides complementary affective cues that further enhance stress detection performance.

4.4 Discussion

Our experiments validate the hypothesis that integrating synthesized visuals with text enhances

stress detection on social media. The significant performance improvement observed with full fine-tuning suggests that adapting the multi-modal embeddings to the domain-specific nuances of stress-related content is critical. Moreover, the ablation study confirms that although text-only models perform strongly, the incorporation of visual cues further improves the detection of subtle affective signals. These findings underscore the potential of multi-modal data fusion for advancing mental health monitoring applications. We hypothesize that the generated visuals act as implicit emotion amplifiers, translating latent affective states into more explicit signals that the model can learn from. The shared embedding space enables the model to reinforce weak cues in one modality using complementary information from the other, thereby improving the robustness of stress detection. While this method shows strong results on the Dreaddit dataset, its generalizability to other mental health tasks or platforms—such as Twitter or Instagram—remains an open question. Future work should explore how this approach adapts to different linguistic styles, content structures, and user populations across platforms.

5 Conclusion

In this work, we introduced a novel multi-modal framework for stress detection using both textual content and synthesized visuals from DALL-E. Leveraging the CLIP model’s robust joint embedding capabilities, our method captures subtle emotional cues missed by text-only approaches. Tested on the Dreaddit dataset, our model achieved 94.90% accuracy with classifier-only training, while full fine-tuning increased performance to 98.41%. These results highlight the significant potential of combining generative image synthesis with multi-modal representation learning for affective computing and mental health monitoring.

Limitations

Despite the promising results of our multi-modal framework, several limitations remain. First, our approach relies on synthesized images generated by DALL-E, which may introduce biases or inconsistencies; the quality and representativeness of the generated visuals can vary depending on the input text. Second, our experiments have been conducted solely on the Dreaddit dataset, and it is unclear whether the observed performance improvements

will generalize to other social media platforms or stress-related domains (Cohan et al., 2018; Mauriello et al., 2021; Garg et al., 2022; Sathvik and Garg, 2023; Chim et al., 2024). Furthermore, while results on the Dreaddit dataset are promising, further research is needed to determine the generalizability of our model across different social media platforms and diverse demographic groups. Finally, even though the fusion of text and visuals enhances stress detection, the interpretability (Jeon et al., 2024) of the resulting multi-modal representations remains an open challenge. Future work should focus on addressing these limitations by exploring more robust image synthesis techniques and developing methods to improve the transparency and interpretability of multi-modal models. One limitation is the lack of systematic evaluation of the generated images. We do not assess whether they reflect the intended affective state or which visual features (e.g., color, composition, expressions) contribute to stress detection. Future work will examine prompt design and affective feature attribution.

Ethical Considerations

Our work involves the analysis of social media data for stress detection, raising important ethical considerations. The use of such data requires strict adherence to privacy protocols and the anonymization of user information. Additionally, generative models like DALL-E can inadvertently propagate biases present in their training data, potentially affecting the fairness and reliability of our system. Care must be taken to ensure that the technology is not misused for surveillance or discriminatory practices. We advocate for responsible usage, transparent reporting of model decisions, and the integration of fairness-aware techniques in future work. As our study uses only anonymized Dreaddit data without new collection or user interaction, ethics approval was not required. Still, using DALL-E to generate images from user content raises concerns. We take precautions against misuse, but future work should pursue consent-driven, transparent frameworks for generative modeling in mental health.

Acknowledgement

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

- Fatma S Abousaleh, Wen-Huang Cheng, Neng-Hao Yu, and Yu Tsao. 2020. Multimodal deep learning framework for image popularity prediction on social media. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):679–692.
- Maryam Mohammed Aldarwish and Hafiz Farooq Ahmad. 2017. Predicting depression levels using social media posts. In *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*, pages 277–280. IEEE.
- Saeed Amal, Lida Safarnejad, Jesutofunmi A Omiye, Iliès Ghanzouri, John Hanson Cabot, and Elsie Gyang Ross. 2022. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in cardiovascular medicine*, 9:840262.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.
- Saurabh Biswas and Yasha Hasija. 2022. Predicting depression through social media. In *Predictive Analytics of Psychological Disorders in Healthcare: Data Analytics on Psychological Disorders*, pages 109–127. Springer.
- Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the clpsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190.
- Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.
- Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Bin Feng, Shulan Ruan, Mingzheng Yang, Dongxuan Han, Huijie Liu, Kai Zhang, and Qi Liu. 2025. Sentiformer: Metadata enhanced transformer for image sentiment analysis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Muskan Garg, Chandni Saxena, Veena Krishnan, Ruchi Joshi, Sriparna Saha, Vijay Mago, and Bonnie J Dorr. 2022. Cams: An annotated corpus for causal analysis of mental health issues in social media posts. *arXiv preprint arXiv:2207.04674*.
- Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. 2023. I can’t believe there’s no images! learning visual tasks using only language supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2672–2683.
- Cosimo Ieracitano, Nadia Mammone, Amir Hussain, and Francesco C Morabito. 2020. A novel multimodal machine learning based approach for automatic classification of eeg recordings in dementia. *Neural Networks*, 123:176–190.
- Hyolim Jeon, Dongje Yoo, Daeun Lee, Sejung Son, Seungbae Kim, and Jinyoung Han. 2024. A dual-prompting for interpretable mental health language models. *arXiv preprint arXiv:2402.14854*.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mental-BERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.
- Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964.
- Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1667–1675.
- Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. 2021. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7.

- Namita Mittal, Divya Sharma, and Manju Lata Joshi. 2018. Image sentiment analysis using deep learning. In *2018 IEEE/WIC/ACM international conference on web intelligence (WI)*, pages 684–687. IEEE.
- Hussein Mouzannar, Yara Rizk, and Mariette Awad. 2018. Damage identification in social media posts using multimodal deep learning. In *ISCRAM*. Rochester, NY, USA.
- Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1505–1514.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.
- Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. 2017. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*.
- MSVPJ Sathvik and Muskan Garg. 2023. Multiwd: Multiple wellness dimensions in social media posts. *Authorea Preprints*.
- Jiayu Shi, Zexiao Wang, Jiandong Zhou, Chengyu Liu, Poly ZH Sun, Erying Zhao, and Lei Lu. 2024. Mentalqlm: A lightweight large language model for mental healthcare based on instruction tuning and dual lora modules. *medRxiv*, pages 2024–12.
- Binyang Song, Rui Zhou, and Faez Ahmed. 2024. Multi-modal machine learning in engineering design: A review and future directions. *Journal of Computing and Information Science in Engineering*, 24(1):010801.
- Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928.
- Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.
- Junyang Wang, Yuanhong Xu, Juhua Hu, Ming Yan, Jitao Sang, and Qi Qian. 2023. Improved visual fine-tuning with natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11899–11909.
- Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.
- Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.
- Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentalama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.
- Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 29.
- Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multimodal data and machine learning techniques: A tutorial and review. *Information fusion*, 59:103–126.
- Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. 2021. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1834.
- Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. 2024. Towards understanding convergence and generalization of adamw. *IEEE transactions on pattern analysis and machine intelligence*.
- Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. 2023. Shifted diffusion for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10157–10166.
- Zhouan Zhu, Shangfei Wang, Yuxin Wang, and Jiaqiang Wu. 2025. Integrating visual modalities with large language models for mental health support. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8939–8954.

A Appendix

In the appendix, we provide further details regarding our experimental setup, hyperparameter settings, and examples of synthesized images. These supplementary materials aim to enhance the reproducibility of our work and offer deeper insights into the performance of our multi-modal framework.

A.1 Hyperparameters for Generating Images with DALL-E

As shown in Table 3, we employed the DALL-E 3 model to synthesize images from social media posts. Our prompt was carefully designed to ensure that the generated visuals consistently capture stress-related cues. For each post, the prompt instructs DALL-E 3 to produce a **consistent, structured** image that visually represents a state of stress or anxiety. This image is expected to include a tense or overwhelming environment (e.g., dim lighting, clutter, urban stress), facial expressions that convey worry, exhaustion, or distress (when humans are depicted), and a darker, cooler color palette to evoke a stressed mood. The images are generated at a resolution of 1024x1024 with standard quality, and one image is produced per post.

A.2 Hyperparameters and Training Setup for CLIP

Table 4 summarizes the hyperparameters and training configurations used in our experiments for both the CLIP Classifier-Only Training and the CLIP Full Fine-Tuning approaches.

In our experiments, the CLIP Classifier-Only Training approach involves freezing the CLIP base and training only the classifier with the AdamW optimizer (Zhou et al., 2024) at a learning rate of 5×10^{-4} , a weight decay of 1×10^{-4} , and a StepLR scheduler (step size of 5 epochs and $\gamma = 0.5$). Training is conducted for up to 10 epochs with early stopping after 3 epochs of no improvement. For the CLIP Full Fine-Tuning approach, both the CLIP base and the classifier are updated. We employ a dual learning rate strategy where the CLIP parameters are optimized at 5×10^{-6} and the classifier at 5×10^{-4} , using the same weight decay and scheduler settings. This configuration runs for up to 15 epochs, with gradient clipping (max norm = 1.0) applied to stabilize training. These hyperparameter choices enable a balanced adaptation of the pre-trained CLIP representations while effectively

learning task-specific features for stress detection.

A.3 Illustrative Examples of Synthesized Visuals from Social Media Posts

In this section, we generate images from social media posts using DALL-E. We provide examples from the Dreddit dataset alongside their corresponding synthesized images (see Figure 2). Each image is generated based on the text of the post, capturing the key emotional and contextual cues embedded within the content. Our approach translates linguistic elements—such as tone, word choice, and contextual details—into visual features, including the color palette, environmental cues, and facial expressions that are indicative of stress. By presenting these paired examples, we illustrate how our multi-modal framework leverages both textual and visual modalities to enhance stress detection, offering a more comprehensive perspective on the underlying affective signals present in social media data.

A.4 Code Availability

The source code for all experiments, including data preprocessing, model training, and evaluation scripts, is available on GitHub: <https://github.com/Efstathia-Soufleri/Stress-Detection-CLIP>. This repository is designed to facilitate the reproducibility of our results and to support further research in this field.

Parameter	Value / Description
Model	dall-e-3
Prompt	Based on the text "{post}", generate a consistent, structured image that visually represents a state of stress or anxiety. The image must include: <ul style="list-style-type: none"> • A tense or overwhelming environment (e.g., dim lighting, clutter, urban stress). • Facial expressions showing worry, exhaustion, or distress (if humans are depicted). • A darker, cooler color palette to evoke a stressed mood.
Size	1024x1024
Quality	Standard
Number of Images	1

Table 3: Summary of DALL-E 3 image generation parameters and prompt design used for synthesizing visuals that capture stress-related cues.

Parameter	Classifier-Only Training	Full Fine-Tuning
Epochs	10	15
Batch Size	32	32
Optimizer	AdamW (classifier only)	AdamW (dual groups)
Learning Rate (Classifier)	5×10^{-4}	5×10^{-4}
Learning Rate (CLIP Base)	—	5×10^{-6}
Weight Decay	1×10^{-4}	1×10^{-4}
LR Scheduler	StepLR (step=5, $\gamma=0.5$)	StepLR (step=5, $\gamma=0.5$)
Early Stopping Patience	3 epochs	3 epochs
Additional Techniques	—	Gradient Clipping (max norm = 1.0)

Table 4: Hyperparameters and training configurations for Classifier-Only Training and Full Fine-Tuning of our proposal.



<p>Image</p> 	<p>Post</p> <p>Especially the power of healing brought upon by service animals. I too, have a service dog named Luna. This wonderful man was nice enough to bring the book back in while I was off yesterday with a note with his name and number telling me to call him when I finish the book. This just made my day, it really did. There's so much negativity in the world today and it seems not many people will stop to do something nice for someone or help them by doing a random act of kindness.</p>
<p>Image</p> 	<p>Post</p> <p>These past couple of months have been the worst. My anxiety has gotten so bad it's effecting my sleep and relationship. I've become so paranoid about my health as well. I don't feel like me anymore and I just feel scared all the time now over every little thing. I don't have money to see a therapist either...</p>

Figure 2: Illustrative examples from the Dreaddit dataset. A social media post and the corresponding synthesized image generated from the post text. These examples demonstrate how our multi-modal framework leverages both textual and visual modalities to capture emotional and contextual cues for enhanced stress detection. The top image and post pair indicate absence of stress and the below pair indicate stress.