CAN VISION MODELS PROCESS PHYSIOLOGICAL SIGNALS? EXPLORING VISUAL TOKENIZATION AS A REPRESENTATION INTERFACE

Frida M. E. Westby* **Li Meng*** **Anis Yazidi*** **Ali Ramezani-Kebrya***
{fmwestby, limeng, anisy, ali}@uio.no

ABSTRACT

Multimodal foundation models increasingly use unified representation interfaces, but the tradeoffs between architectural unification and domain-specific tokenization remain unclear. We explore visual tokenization as a representation interface for emotion recognition by converting physiological signals and text into images processed by frozen Vision Transformers. This parameter-efficient approach, which trains only 0.85% of weights, enables multimodal learning without modality-specific encoders. We identify when frozen pretrained features succeed and when they require domain adaptation. We also discuss design considerations for balancing unified processing with domain-appropriate inductive biases.

1 INTRODUCTION

Recent advances in multimodal foundation models have led to the adoption of unified visual tokenizations that process diverse inputs using shared architectures (Team, 2025; Zhou et al., 2024). However, important questions remain regarding the circumstances under which architectural unification outperforms specialized designs and the effectiveness of transferring frozen pretrained representations across domains.

We investigate these questions using physiological emotion recognition, a domain in which traditional methods depend on extensive feature engineering (Greco et al., 2016).The proposed MIMI (Multi-modal Image-based Emotion Recognition) framework transforms physiological signals, such as electrodermal activity (EDA) and photoplethysmography (PPG), along with text descriptions, into RGB images. These images are then processed by frozen Vision Transformers pretrained on ImageNet (Deng et al., 2009).

The primary contributions are as follows: (1) Introduction of visual tokenization as a parameter-efficient interface for heterogeneous physiological and textual data; (2) Analysis of the domain alignment challenges encountered when repurposing pretrained visual features for non-visual modalities; and (3) Discussion of design considerations for balancing unified tokenization with domain-appropriate inductive biases.

2 METHODS

We combine physiological and textual data into a single visual token format. In the following sections, we explain the visual tokenization process, the model components, and the experimental procedures.

2.1 VISUAL TOKENIZATION PIPELINE

Signal-to-Image Rendering: Physiological signals are converted into visual representations using Short-Time Fourier Transforms (STFT). Electrodermal activity (EDA, 15.625 Hz) and photoplethysmography (PPG, 125 Hz) signals are transformed using a 256-sample Hann window with 50% over-

*University of Oslo, Norwegian Centre for Knowledge-driven Machine Learning (Integreat), TRUST – The Norwegian Centre for Trustworthy AI, MishMash Centre for AI and Creativity

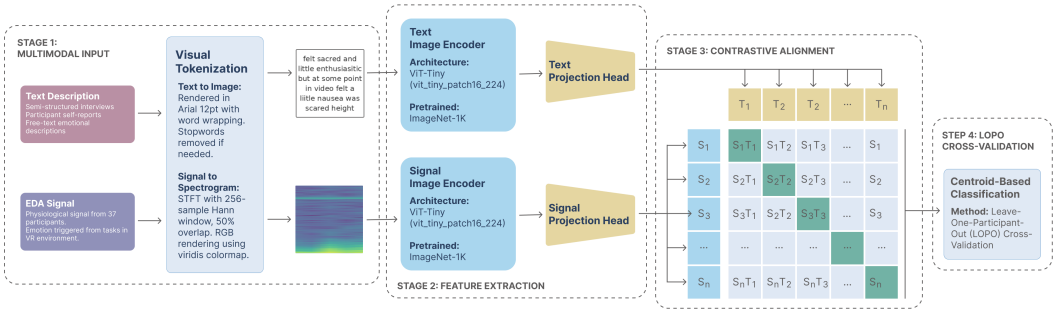

Figure 1: MIMI pipeline: Multimodal inputs are converted to RGB images, processed via frozen ImageNet-pretrained ViTs, and aligned using NT-Xent loss. Evaluation uses LOPO cross-validation.

lap. This fixed window size results in different temporal spans for each modality: 16.4 seconds for EDA and 2.0 seconds for PPG.

To manage the high degree of inter-participant variability in physiological data, spectrogram intensities are normalized to the 5th-95th percentile range, which helps mitigate the influence of outliers. These normalized spectrograms are then rendered as RGB images using the viridis colormap. We chose this colormap for its perceptual uniformity, as it provides consistent intensity gradations that may help the frozen Vision Transformer distinguish subtle frequency variations that appear as hue changes in the rendered image. This rendering process ensures the signals are compatible with the three-channel architecture of ImageNet-pretrained models.

Text-to-Image Rendering: Participant descriptions are rendered as images (Arial 12pt) with stopword removal to manage text length. This approach converts text into images, enabling us to investigate whether a single frozen Vision Transformer can process both text and physiological signals.

Unified Visual Processing: All images (224×224, ImageNet normalized) are processed through frozen ViT-Tiny backbones (5.9M parameters) pretrained on ImageNet-1K (Deng et al., 2009; Dosovitskiy et al., 2021). We extract 192-dimensional features from the classification token, which aggregates information from all image patches, while keeping all backbone parameters frozen.

Contrastive Alignment: Each modality has a trainable projection head (Dropout p=0.3, Linear 192→128, L2-norm, approximately 25K parameters) that maps ViT features to a shared 128-dimensional embedding space. Training is performed using NT-Xent loss (Chen et al., 2020; Oord et al., 2019) with a temperature of $\tau = 0.07$.

2.2 Experimental Setup

The approach is evaluated on the EEVR dataset Singh et al. (2024), which contains physiological recordings from 37 participants across 296 emotion-elicitation tasks, totaling 797 minutes of data. The dataset covers Russell's circumplex model of affect Russell (1980), with videos spanning High/Low Valence and High/Low Arousal quadrants. Three binary classification tasks are defined based on Self-Assessment Manikin (SAM, (Bradley & Lang, 1994)) ratings: Arousal classification distinguishes low (scores 1-3) from high (scores 4-5) activation levels; Valence classification separates negative (1-3) from positive (4-5) affective quality; and Task-wise engagement classifies high-engagement videos (HVHA/HVLA) against lower-engagement content and baseline recordings.

Training is performed using the Adam optimizer Kingma & Ba (2017) with a learning rate of 5×10^{-4} and weight decay of 0.01. The learning rate schedule consists of a 2-epoch linear warmup, followed by cosine annealing to 1×10^{-6} over 20 total epochs. A batch size of 16 is used, and all experiments are repeated with three random seeds (42, 43, 111) to assess stability and statistical reliability.

We perform leave-one-participant-out (LOPO, based on leave-one-subject-out shown by Saganowski et al. (2023)) cross-validation by systematically holding out each of the 37 participants, training on the remaining 36, and evaluating on the held-out participant's data. This procedure en-

sures participant-independent evaluation, critical for assessing generalization to unseen individuals. At test time, we use only the signal modality for classification, applying centroid-based assignment where test embeddings are assigned to the nearest class centroid in the learned embedding space.

3 Results and Discussion

The experimental results establish the feasibility of visual tokenization and elucidate the trade-offs between architectural unification and domain-specific design.

3.1 Parameter Efficiency and Task-Dependent Alignment

MIMI optimizes a limited subset of parameters using lightweight projection heads while maintaining a frozen visual backbone. This parameter-efficient architecture facilitates edge deployment, supports multi-task systems with shared backbones, and enables rapid adaptation to new tasks.

We evaluated MIMI against a similar CLSP baseline as presented in (Singh et al., 2024). Unlike Singh et al. (2024), we use raw data rather than feature-engineered data. Our evaluation, conducted across three binary classification tasks using leave-one-participant-out (LOPO) cross-validation, shows that MIMI achieves consistent improvements over baseline, as detailed in Table 1.

Signal	Method	Task-wise Acc / F1	Valence Acc / F1	Arousal Acc / F1
EDA	MIMI	0.527 / **0.561**	0.537 / **0.621**	0.488 / **0.558**
	CLSP	0.506 / 0.513	0.512 / 0.512	0.447 / 0.488
PPG	MIMI	0.531 / **0.567**	0.546 / **0.634**	0.485 / **0.554**
	CLSP	0.509 / 0.519	0.512 / 0.512	0.451 / 0.488
Improvement (MIMI - CLSP)				
EDA	Δ Acc	+0.021	+0.025	+0.041
	Δ F1	+0.048	**+0.109**	+0.070
PPG	Δ Acc	+0.022	+0.034	+0.034
	Δ F1	+0.048	**+0.122**	+0.066

Table 1: Performance comparison between MIMI and CLSP baseline across EDA and PPG signals. Values show mean ± standard deviation over 111 leave-one-participant-out folds (37 participants × 3 seeds). MIMI shows consistent improvements across all metrics, with strongest gains for valence F1 (+10.9% EDA, +12.2% PPG).

Approach	Signal	Task-wise	Valence	Arousal
Feature-based (EEVR)	EDA	**90.8%**	61.9%	57.2%
	PPG	81.0%	61.3%	56.3%
MIMI (This work)	EDA	52.7%	53.7%	48.8%
	PPG	53.1%	54.6%	48.5%
CLSP (This work)	EDA	50.6%	51.2%	44.7%
	PPG	50.9%	51.2%	45.1%
Gap (Feature - MIMI)				
	EDA	-38.1 pp	-8.2 pp	-8.4 pp
	PPG	-27.9 pp	-6.7 pp	-7.8 pp

pp = percentage points

Table 2: Accuracy comparison (%) between feature-based and spectrogram-based approaches. Negative "Gap" values indicate the percentage point (pp) deficit of MIMI compared to feature-engineered methods.

While these improvements are consistent, paired t-tests reveal that most gains do not survive the Bonferroni-corrected significance threshold of $\alpha^* = 0.05/6 \approx 0.0083$ Dunn (1961). The primary exception is Valence F1 ($p = 0.009$, Cohen's $d = 0.539$), which demonstrates a medium effect size and a 95% bootstrap confidence interval that excludes zero ($[+0.041, +0.178]$) Cohen (2013). This suggests that while visual tokenization provides a flexible representation interface, the practical significance of the gains is most pronounced for tasks with high semantic grounding, such as emotional valence.

Alignment quality is strongly influenced by task semantics. Valence, defined as the pleasantness or unpleasantness dimension (Averill, 1975), benefits substantially from contrastive text-signal alignment (+10.9% F1), whereas arousal (Nowlis, 1965) and engagement demonstrate more modest improvements (+5–7% F1). This trend indicates a relationship between linguistic expressibility and alignment effectiveness: when affective states are easily verbalized (for example, "I felt happy"), text-based supervision is more effective. Conversely, arousal and engagement, although physiologically measurable, are less consistently expressed in language, which may constrain the effectiveness of text-based alignment (Fazzi et al., 2025; Mendes & Martins, 2023; Russell, 1980).

These preliminary findings indicate that unified visual tokenization can achieve resource efficiency; however, its effectiveness is contingent upon semantic expressibility. Language supervision is most advantageous for concepts that are naturally verbalized. For experiential dimensions that are less

accessible through language, alternative supervision strategies such as demonstrations, preference comparisons, or direct sensor feedback may be required.

3.2 Domain Mismatch and Transfer Learning Limitations

ImageNet features, which are optimized for natural images such as edges, textures, and objects, do not transfer effectively to physiological spectrograms. These spectrograms are defined by frequency bands and temporal evolution (Deng et al., 2009). The horizontal frequency stratification in spectrograms encodes physiological dynamics, which is fundamentally distinct from the spatial correlations present in natural images.

Electrodermal activity (EDA) and photoplethysmography (PPG) signals demonstrate similar performance patterns despite their distinct physiological origins. This finding supports the use of unified processing approaches but also highlights inherent limitations in transferability. While architectural unification treats both modalities equivalently, it fails to capture domain-specific structural characteristics.

Furthermore, transferability challenges also affect our text-to-image rendering strategy. The frozen ViT backbone was pretrained on ImageNet to recognize natural objects and textures, so it is not optimized for OCR-like character recognition or extracting fine-grained linguistic features. As a result, the model may rely on superficial visual patterns, such as the length of the rendered text block or spatial density, instead of a deep semantic understanding of the emotional descriptions. This offers important empirical intuition for the observed performance ceiling: while visual tokenization provides a unified architectural interface, it likely discards the rich structural and semantic priors found in traditional text embeddings.

3.3 Data Scale and the Inductive Bias Trade-off

With limited training data, such as in the EEVR dataset (Singh et al., 2024), unified visual tokenization cannot fully overcome the mismatch between pretrained ImageNet features and physiological spectrograms. As noted in Section 3.1, while MIMI provides consistent gains over the CLSP baseline, it remains significantly below the performance of traditional pipelines.

Feature-based methods encoding explicit physiological knowledge substantially outperform our unified approach by injecting decades of domain expertise into the feature space. To quantify this performance gap, we compare MIMI against the feature-engineered results reported in the original EEVR study (Singh et al., 2024).

As shown in Table 2, there is a big 38.1 percentage point gap in task-wise accuracy for EDA signals. This shows that while visual tokenization reduces engineering effort and offers a flexible interface, domain-specific inductive biases remain critical in data-scarce settings. With only 37 participants, the frozen ViT backbone lacks enough data to learn robust physiological mappings that hand-crafted features such as SCR peaks or HRV metrics provide out-of-the-box.

This analysis highlights a fundamental trade-off. Unified approaches reduce engineering effort but may forgo domain-specific inductive biases that are critical in data-scarce settings. In data-rich environments, unified architectures can learn these patterns. In contrast, when data is limited, explicit knowledge integration becomes essential.

These findings imply that architectural decisions should be data-dependent. High-resource domains benefit from unified end-to-end learning, whereas low-resource or specialized domains require hybrid approaches that combine frozen encoders with domain-specific pathways.

3.4 Hypoteses for multimodal design

The results of this study provide a basis for the following hypotheses regarding the design of multimodal architectures in specialized domains. Rather than selecting exclusively between unified processing and domain-specific engineering, hybrid architectures that combine frozen, pretrained encoders with lightweight, domain-specific pathways should be considered. Importantly, this work is an interface and transferability study rather than a claim that generic vision models are a state-

of-the-art replacement for specialized physiological encoders, which still have a significant performance advantage in low-data regimes.

The choice of visual tokenization should be guided by the specific characteristics of the domain. Visual tokenization is most effective when three criteria are met: (1) the domain contains exploitable visual structure, (2) language provides meaningful supervision, and (3) parameter efficiency is prioritized. If these conditions are not met, alternative strategies such as learned embedding spaces or domain-specific architectural components should be explored.

To address the performance gap observed in this study, we propose an unvalidated hypothesis for future investigation: a three-stage pretraining approach. This proposed framework includes (1) broad visual pretraining (e.g., ImageNet), (2) intermediate domain-adaptive pretraining on relevant corpora such as diverse physiological spectrograms, and (3) task-specific fine-tuning. While this staged approach aims to balance transfer efficiency with domain specificity, but it remains unvalidated in the current work and serves as a roadmap for future empirical testing.

Evaluation protocols should incorporate small-data scenarios to reveal architectural inductive biases. While large-scale benchmarks assess model capacity, small-data performance is essential for real-world deployment, where labeled data is often costly or restricted by privacy constraints.

4 Conclusion

We investigated visual tokenization as a case study in visual tokenization design for multimodal emotion recognition, training only 0.85% of the parameters by employing frozen Vision Transformers. Our analysis identifies three conditions that determine the success of frozen transfer: semantic grounding, where tasks with linguistic correlates benefit from language supervision; domain alignment, which requires pretraining distributions to match deployment requirements; and data scale, as limited samples increase the importance of inductive bias. These findings indicate that unified architectures exchange domain-specific knowledge for architectural simplicity, with this trade-off contingent on data availability and deployment constraints.

Our findings prompt broader questions for the development of multimodal foundation models: (1) Under what circumstances is cross-domain tokenization, such as mapping non-visual to visual representations, preferable to modality-specific encoders? (2) How should the simplicity of unified architectures be balanced against the advantages of domain-specific inductive biases? (3) Can intermediate pretraining stages effectively bridge the gap between general visual features and specialized domains? We encourage the research community to contribute perspectives on these tokenization tradeoffs.

Future research should investigate hybrid architectures that integrate frozen encoders with domain-specific pathways, utilize intermediate pretraining to address distribution gaps, and develop parameter-efficient adaptation methods. Our study clarifies the conditions under which architectural unification benefits multimodal learning and when specialized expertise remains necessary.

Acknowledgments

This work was supported by the Research Council of Norway through FRIPRO Grant under project number 356103 and its Centres of Excellence scheme, Integreat - Norwegian Centre for knowledge-driven machine learning under project number 332645. The computations were performed on resources provided by Educloud Research infrastructure at UiO.

REFERENCES

James R Averill. *A semantic atlas of emotional concepts*. 1975.

Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, March 1994. ISSN 00057916. doi: 10.1016/0005-7916(94)90063-9. URL https://linkinghub.elsevier.com/retrieve/pii/0005791694900639.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, July 2020. URL http://arxiv.org/abs/2002.05709. arXiv:2002.05709 [cs].

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, May 2013. ISBN 978-1-134-74270-7. Google-Books-ID: 2v9zDAsLvA0C.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/abstract/document/5206848. ISSN: 1063-6919.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961. ISSN 0162-1459. doi: 10.1080/01621459.1961.10482090. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1961.10482090.

Gino Franco Fazzi, Julie Skoven Hinge, Stefan Heinrich, and Paolo Burelli. Don't Get Too Excited – Eliciting Emotions in LLMs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9, April 2025. doi: 10.1145/3706599.3720191. URL http://arxiv.org/abs/2503.02457. arXiv:2503.02457 [cs].

Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE transactions on bio-medical engineering*, 63(4):797–804, April 2016. ISSN 1558-2531. doi: 10.1109/TBME.2015.2474131.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Gonçalo Azevedo Mendes and Bruno Martins. Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers, February 2023. URL http://arxiv.org/abs/2302.14021. arXiv:2302.14021 [cs].

Vincent Nowlis. Research with the Mood Adjective Check List. In *Affect, cognition, and personality: Empirical studies*, pp. xii, 464–xii, 464. Springer, Oxford, England, 1965.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980. ISSN 1939-1315. doi: 10.1037/h0077714.

Stanisław Saganowski, Bartosz Perz, Adam G. Polak, and Przemysław Kazienko. Emotion Recognition for Everyday Life Using Physiological Signals From Wearables: A Systematic Literature Review. *IEEE Transactions on Affective Computing*, 14(3):1876–1897, July 2023. ISSN 1949-3045. doi: 10.1109/TAFFC.2022.3176135. URL https://ieeexplore.ieee.org/document/9779458.

Pragya Singh, Ritvik Budhiraja, Ankush Gupta, Anshul Goswami, Mohan Kumar, and Pushpendra Singh. EEVR: A Dataset of Paired Physiological Signals and Textual Descriptions for Joint Emotion Representation Learning. November 2024. URL https://openreview.net/forum?id=qgzdGyQcDt#discussion.

Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models, March 2025. URL http://arxiv.org/abs/2405.09818. arXiv:2405.09818 [cs].

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, August 2024. URL http://arxiv.org/abs/2408.11039. arXiv:2408.11039 [cs].

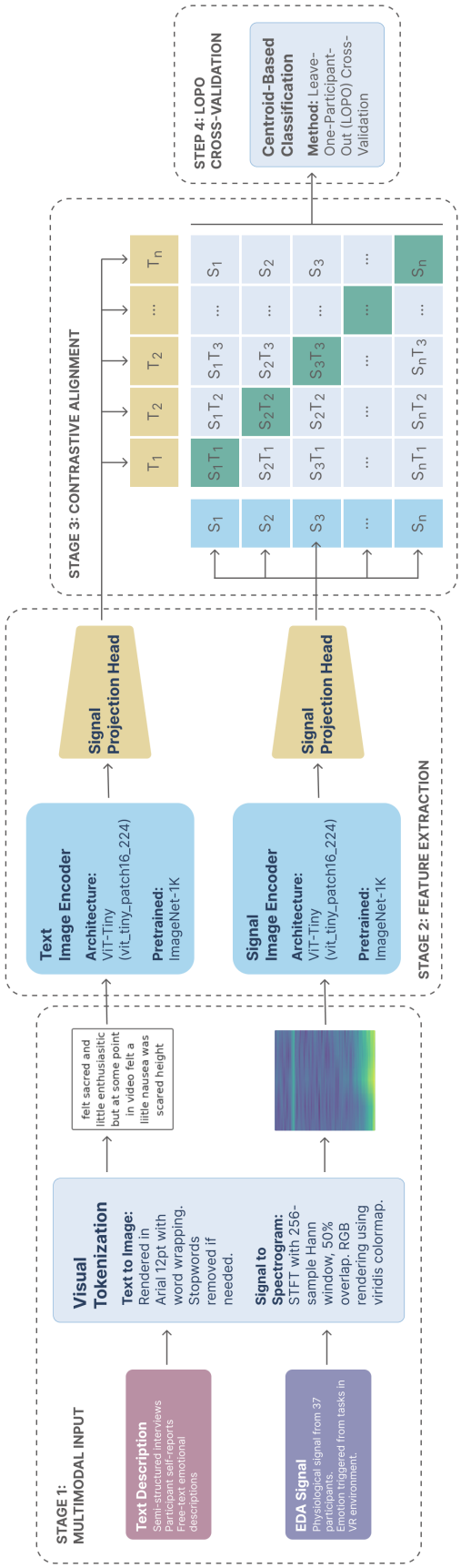

Figure 2: **MIMI pipeline:** Raw EDA and PPG signals and participant text descriptions are converted into RGB images, spectrograms and rendered text, respectively. Both modalities are processed through frozen ViT-Tiny encoders pretrained on ImageNet-1K as shown in blue (Dosovitskiy et al., 2021; Krizhevsky et al., 2012). Small trainable projection heads map the 192-dimensional ViT features to 128-dimensional embeddings. The NT-Xent contrastive loss aligns these embeddings during training, while test-time classification uses only the signal modality with centroid-based prediction.