# Interpretable Concept-based Deep Learning Framework for Multimodal Human Behavior Modeling

Xinyu Li, Marwa Mahmoud

[a] *School of Computing Science, University of Glasgow, Glasgow, G12 8RZ, United Kingdom*

## 1. Expanded Discussion of AGCM

The use of handcrafted features, such as AU detections, has been ongoing for decades. These approaches mainly focus on automatically mapping the facial representation to a single numerical value, without fully accounting for the complexity of one's affective state. Like in most of the feature-based approaches, relying solely on these numerical values for intricate AC tasks risks overlooking other emotion-related information conveyed by the subject, potentially degrading performance. Similarly, in multi-task learning—for instance, simultaneously predicting AU and expression—each classification head optimizes independently, rather than fostering mutually beneficial learning that emphasizes the relevance of AUs to facial expressions.

In contrast, as illustrated in Fig. 1, the proposed AGCM framework enhances both model explainability and performance by bridging this gap. It employs an end-to-end learning strategy that quantifies the contributions of underlying emotion-related indicators to the final task prediction. By design, AGCM naturally advances traditional feature-based and multi-task AC approaches, where feature representations are either static or insufficient as explanations for task predictions.

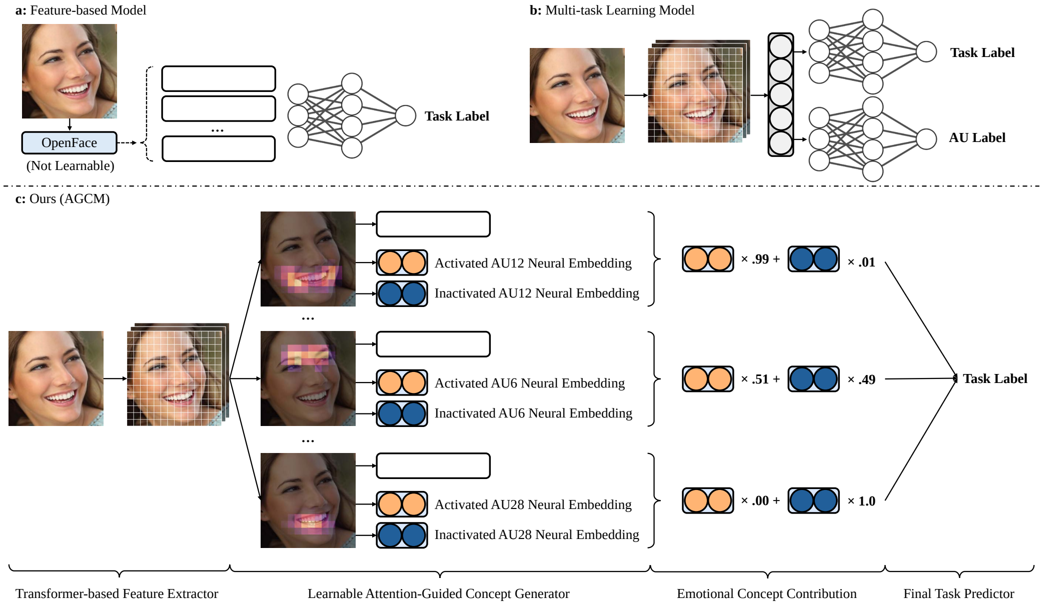
a: Feature-based Model
OpenFace
(Not Learnable)
...
Task Label
b: Multi-task Learning Model
Task Label
AU Label
c: Ours (AGCM)
Activated AU12 Neural Embedding
Inactivated AU12 Neural Embedding
× .99 +
× .01
...
Activated AU6 Neural Embedding
Inactivated AU6 Neural Embedding
× .51 +
× .49
Task Label
...
Activated AU28 Neural Embedding
Inactivated AU28 Neural Embedding
× .00 +
× 1.0
Transformer-based Feature Extractor
Learnable Attention-Guided Concept Generator
Emotional Concept Contribution
Final Task Predictor

Figure 1: (a) Feature-based models rely on manual feature preprocessing using external automatic toolkits, such as OpenFace, which operate outside the model's training loop and are not learnable. These models map preprocessed features to task labels, risking the loss of valuable raw data information that could contribute to more comprehensive predictions. (b) Multi-task learning models train multiple tasks independently, with the learning of specific emotional tasks and AUs being uncorrelated and disconnected. As a result, AU predictions in multi-task learning cannot effectively explain the emotional predictions, limiting the interpretability of the model. (c): The proposed AGCM framework operates as follows: after feature extraction, the Attention-Guided Concept Generator creates learnable neural representations for both activated and inactivated concepts, along with their respective activation scores. It then computes the emotional concept contribution by combining the activated and inactivated embeddings for each concept. Parameter optimization for concept learning is conducted concurrently with task-label learning in an end-to-end manner, enabling the model to capture emotional concept contributions while effectively overcoming the trade-off between explainability and performance.

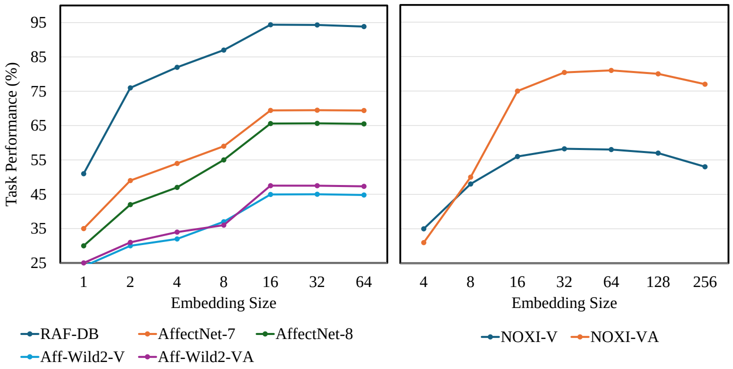

Figure 2: Task performance evaluation (%) with different embedding sizes. For RAF-DB and AffectNet, the overall accuracy is reported. For Aff-Wild2 and NOXI, the F-1 score and CCC score are reported.

## 2. Embedding Size Ablation Study

Previous studies have demonstrated that embedding size can impact the task performance of concept-based frameworks [1]. The optimal concept size may vary depending on the task. In this work, we use an embedding size of 16 for all FER tasks and 32 for engagement estimation tasks.

Fig. 2 shows the task performance across various embedding sizes. For both applications, performance initially improves with increasing embedding size. However, once the embedding size reaches the limitation of the model's learning capacity, further increases do not yield performance gains. Instead, larger embeddings may significantly raise the number of parameters, which can pose challenges for model training and deployment.

## 3. Comparing End-to-end and By-step AGCM

To further assess the efficiency of the AGCM framework, we compare the end-to-end and by-step training strategies. In by-step AGCM, the model first optimizes a mapping function from the raw input to all intermediate concept scores.

Table 1: Performance comparison (%) of the end-to-end and by-step AGCM framework. For RAF-DB and AffectNet, the overall accuracy is reported. For Aff-Wild2 and NOXI, the F-1 score and CCC score are reported.

| | Data | End-to-end AGCM | By-step AGCM |
|---|---|---|---|
| RAF-DB | V | **94.40** | 89.71 |
| AffectNet-7 | V | **69.45** | 64.08 |
| AffectNet-8 | V | **65.62** | 61.36 |
| Aff-Wild2 | V | **44.95** | 39.10 |
| Aff-Wild2 | V/A | **47.52** | 39.23 |
| NOXI | V | **59.24** | 52.01 |
| NOXI | V/A | **80.39** | 67.88 |

If the concepts include only AUs, this phase operates similarly to an AU detector, generating activation probabilities for all AUs. These AU probabilities are then combined with the embeddings in a subsequent optimization step to predict the final facial expression label separately.

In by-step AGCM, the neural embeddings of intermediate concepts are not trainable during task learning. The parameter optimization treats the concept and task loss separately. This approach contrasts with end-to-end training, where a unified push-pull joint loss is employed to enhance both concept explainability and task performance simultaneously.

Table 1 presents a performance comparison between the end-to-end and by-step AGCM training strategies. Compared to the end-to-end approach, the by-step training strategy results in performance degradation across all datasets, with particularly notable declines in the multimodal AGCM framework, where separately learning concepts can lead to significant information loss from the raw data. Thus, we posit that jointly learning the concept and task label enhances both model explainability and task performance by compelling the model to explicitly supervise human-understandable features derived from domain-specific prior knowledge.

## 4. Expanded Discussion of AGCM and Map-based XAI

Map-based XAI was originally designed for general ML tasks like object localization, where attention heatmaps serve as effective tools to indicate object locations [2]. In affective signal processing, however, spatial concept explanations offer significant advantages over map-based XAI by providing domain-specific insights alongside task performance improvements. Simply presenting an attention heatmap over a face region offers minimal value for domain experts in AC applications. For instance, two opposing indicators, AU12 (Lip Corner Puller) and AU15 (Lip Corner Depressor), appear in the same region of the face, making it insufficient to rely solely on attention maps for emotion interpretation. Instead, conceptual explanations that explicitly indicate the activation and contribution of specific AUs provide a more natural and informative approach to AC tasks.

Recent map-based FER work [3] uses pre-generated AU maps based on emotion labels to guide model learning, depending on a strict mapping between AUs and facial expressions. For example, for images labeled as “happiness,” this approach restricts the model’s focus strictly to the AU6 and AU12 regions, regardless of whether these specific AUs are activated, ignoring other facial information that may contribute to the expression. This rigid mapping not only degrades performance but also proves limiting in downstream AC applications, such as engagement estimation or mental health assessment, where there is no clear mapping between AUs and affective labels.

Fig. 3 compares explanations provided by our proposed AGCM with those from two map-based XAI methods [2, 3]. The attention heatmaps from the map-based XAI approaches appear similar across different expression labels, offering insufficient interpretability for high-stakes AC applications. In contrast, AGCM

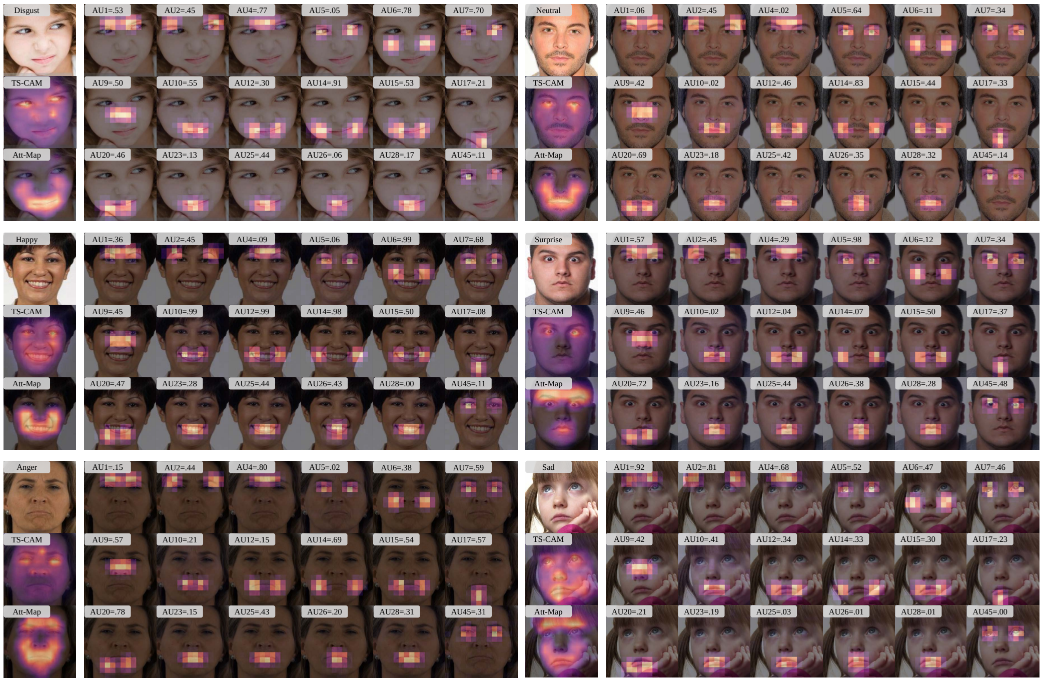

Figure 3: Explanation examples of map-based TS-CAM [2], attention map-based FER (Att-Map) [3], and the proposed AGCM framework. In addition to all concept locations, AGCM explicitly provides the contribution score of each concept, offering domain-specific insight into the model decision-making process. The images are randomly picked from the AffectNet test set.

not only localizes each AU but also quantifies its contribution to the final prediction, delivering richer insights into model predictions while achieving state-of-the-art task performance.

## References

[1] M. E. Zarlenga, B. Pietro, C. Gabriele, M. Giuseppe, F. Giannini, M. Diligenti, S. Zohreh, P. Frederic, S. Melacci, W. Adrian, et al., Concept embedding models: Beyond the accuracy-explainability trade-off, in: Advances in Neural Information Processing Systems, Vol. 35, Curran Associates, Inc., 2022, pp. 21400–21413.

[2] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, Q. Ye, Ts-cam: Token semantic coupled attention map for weakly supervised object localization, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2886–2895.

[3] S. Belharbi, M. Pedersoli, A. L. Koerich, S. Bacon, E. Granger, Guided interpretable facial expression recognition via spatial action unit cues, arXiv preprint arXiv:2402.00281 (2024).