Causally Inspired Attention Supervision for Affective Behaviour Analysis

Nemanja Rašajski[1], Konstantinos Makantasis[2], Antonios Liapis[1], and Georgios N. Yannakakis[1]

[1] Institute of Digital Games, University of Malta, Malta
{nemanja.rasajski,antonios.liapis,georgios.yannakakis}@um.edu.mt
[2] AI Department, University of Malta, Malta
konstantinos.makantasis@um.edu.mt

Abstract. Affective Behaviour Analysis enables machines to infer human affective states from behavioural signals, particularly facial expressions, in real-world environments. This remains a challenging problem because affective cues must be distinguished from confounding factors such as illumination, head pose and demographic variation. To advance research in this area, the *11th Affective Behaviour Analysis in-the-wild Competition* includes the Multi-Task Learning Challenge, in which participants develop a unified framework for Valence-Arousal Estimation, Expression Recognition, and Action Unit Detection. Attention mechanisms have proven effective across a wide range of tasks by learning to aggregate the most informative features for prediction. However, they may still rely on spurious correlations rather than invariant affective cues, resulting in reduced generalization to unseen subjects and environments. To address this limitation, we propose an attention pooling framework that encourages consistent attention across subjects, inspired by causal learning. Specifically, we introduce causal supervision to encourage attention to facial regions with predictive value across subjects. Furthermore, motivated by causal representation learning, we apply an independence regularizer between the Key and Value projections to encourage complementary, non-redundant representations. Our best configuration improves the composite affective multi-task metric by 0.14 over attention pooling, while ablations confirm the necessity of each proposed change. Code is publicly available online.[3].

1 Introduction

Affective computing [48] seeks to develop intelligent systems that perceive and interpret human emotions from signals such as facial expressions, speech, and text, enabling applications in education [52], healthcare [1,17], and robotics [43, 56]. Multi-task learning (MTL) jointly optimizes multiple related tasks through shared representations, enabling knowledge transfer across tasks and improving generalization by exploiting complementary information across supervisory

[3] https://github.com/nrasajski/ABAW11

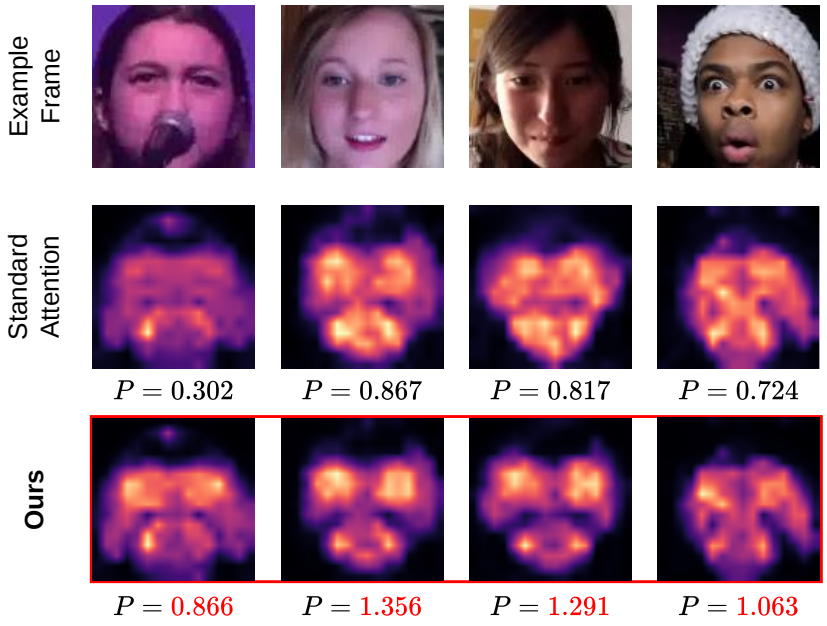

Fig. 1: Attention heatmaps for multi-task prediction on representative frames sampled from four videos in the s-Aff-Wild2 dataset using the FRoundation backbone. The proposed method (red) consistently focuses on discriminative facial regions while suppressing attention to less informative areas. Compared with the baseline using standard attention pooling, our approach achieves a higher video-level multi-task score P, defined as the sum of the task-specific evaluation metrics for Valence-Arousal estimation, Expression Recognition, and Action Unit detection.

signals [14]. MTL is particularly well suited to affective computing, where emotion is commonly characterized using three complementary representations: the Ekman categorical model [7, 13], the Russell circumplex model [16, 51], and the Facial Action Coding System [12]. Modelling these representations as related supervisory tasks enables the joint learning of shared and task-specific information, which has been shown to improve predictive performance [58]. Building on this paradigm, the Affective Behaviour Analysis in-the-Wild (ABAW) Competition introduced the MTL Challenge, in which models simultaneously perform valence-arousal (VA) estimation, expression recognition (EX), and action unit (AU) detection.

However, affective behaviour analysis remains challenging because emotion-relevant features are often confounded by factors such as pose, illumination, and environmental context, which may correlate with emotion labels without being subject invariant. As a result, deep models can rely on these shortcuts rather

than learning emotion-related cues, leading to poor generalization on unseen subjects. This limitation persists even in attention-based models, despite their strong performance across a wide range of tasks [62].

To improve generalization, we propose a causally inspired attention pooling a framework that promotes robust, emotion-relevant cues while reducing the influence of confounding factors through two key ideas. First, causal supervision guides attention towards relatively invariant facial regions using counterfactual attention weights. Second, inspired by causal representation learning [54], we encourage the Key and Value projections to learn mutually orthogonal representations. In the attention mechanism, the Key projection determines how input features are matched to attention queries, while the Value projection encodes the information aggregated by the attention weights [60]. Encouraging these projections to capture distinct information reduces redundancy and promotes more robust representations.

As illustrated in Fig. 1, our approach consistently identifies discriminative facial regions despite variations in identity, illumination, and facial accessories, encouraging the model to attend to emotion-relevant regions while suppressing nuisance factors. Extensive experiments and ablation studies on the s-Aff-Wild2 dataset demonstrate consistent improvements over standard attention pooling. Under a five-fold cross-validation protocol, the proposed method increases the composite score P (defined as the sum of metrics for VA estimation, ER, and AU detection) from 0.87 to 0.95 with the DINOv2 [45] visual encoder and from 1.05 to 1.19 with the FRoundation [9] encoder.

2 Related Work

Aff-Wild2 [19–36, 66] is a widely used benchmark dataset for in-the-wild affective behaviour analysis. It provides frame-level continuous valence-arousal annotations, enabling the modelling of subtle emotional dynamics under naturalistic conditions. The dataset comprises 594 videos (nearly 3 million frames) from 584 subjects with diverse demographic characteristics. In this work, we use the s-Aff-Wild2 subset, which additionally provides labels for the six basic emotions and 12 AU. Together, these annotations support the complementary analysis of affective states and facial behaviour, with VA estimation capturing emotional pleasantness and activation, EX classifying discrete facial expressions, and AU detection identifying the underlying facial muscle movements.

In the context of multi-task learning for affective computing, Xia and Liu [63] propose a framework that jointly learns categorical emotion recognition and continuous VA estimation, demonstrating that shared representations across these related tasks improve emotion recognition performance. Deng et al. [10] introduce a teacher-student framework [18] that jointly predicts AU, categorical expressions, and continuous VA while effectively handling incomplete annotations across heterogeneous emotion datasets. Building on these approaches, Toisoul et al. [58] use facial landmarks to generate attention maps that highlight subtle fa-

cial regions, combining multi-task learning over discrete and continuous emotion labels with knowledge distillation in a teacher–student framework.

When it comes to work on s-Aff-Wild2, Liu et al. [40] first perform self-supervised pre-training on face datasets before adopting a progressive learning strategy that transitions from single-task learning to joint multi-task learning. Savchenko et al. [53] combine multiple facial feature extractors pre-trained in a multi-task setting and further refine predictions through post-processing techniques, including temporal smoothing and prediction blending. Cabacas-Maso et al. [6] employ a Dual-Direction Attention Mixed Feature Network [68] for multi-task facial expression recognition. Li et al. [39] introduce a task-adaptive block that performs cross-attention between a set of learnable query vectors and pre-extracted features to extract task-specific representations, together with an AU-assisted Graph Convolutional Network.

Recent advances in causal representation learning aim to improve robustness and generalization by learning representations that capture task-relevant factors while reducing the influence of spurious correlations and dataset-specific confounders [54]. For example, Wang et al. [61] employ causal interventions to supervise graph attention, encouraging attention weights that more accurately reflect underlying causal relationships. In affective computing, recent work has similarly incorporated causal learning to improve the robustness of affective representations by mitigating the effects of confounding factors. Chen et al. [8] formulate visual emotion recognition as a causal intervention problem to reduce dataset bias, while Yang et al. [64] explicitly model contextual information as a confounder and perform deconfounded emotion recognition.

Our work aims to improve multi-task affect prediction by incorporating causally motivated supervision into an attention-based representation learning framework, rather than explicitly modelling causal relationships through interventions or causal graph inference. Specifically, we build on the causal supervision strategy for attention proposed by Wang et al. [61] and extend it by encouraging complementary Key and Value embeddings, leading to more discriminative shared representations that benefit multiple affective prediction tasks.

3 Method

We build on attentive feature aggregation [3], which computes the similarity between a query (Q) and a set of keys (K), scales the resulting scores, applies a softmax function to obtain attention weights, and uses these weights to compute a weighted sum of the corresponding values (V). The query specifies the information to retrieve, the keys determine the relevance of each feature to the query, and the values contain the information to be aggregated. We introduce three modifications to this formulation. First, we generate counterfactual attention-weight examples and use them as causally motivated supervision to guide attention towards subject-independent, emotion-relevant facial regions. Second, motivated by the Independent Causal Mechanisms principle [54], we regularize the cross-covariance between the K and V representations so that they encode comple-

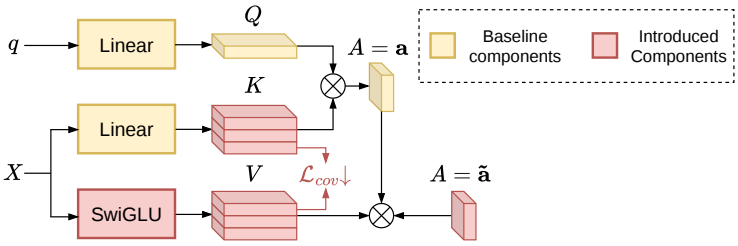

Fig. 2: Our proposed modifications: counterfactual attention supervision, cross-covariance regularization of the K and V representations, and a nonlinear SwiGLU V projection instead of the standard linear projection. All proposed modifications are highlighted in red.

mentary information. Finally, we replace the standard linear V projection with a SwiGLU-based non-linear projection [55] to increase representational capacity. These modifications are illustrated in Fig. 2.

The overall multi-task prediction pipeline consists of four stages. First, facial images are converted into patch embeddings using a frozen pre-trained ViT backbone. Second, our proposed attentive aggregation module pools the patch embeddings into a single frame-level representation. Third, the sequence of frame-level representations is processed by a Temporal Convolutional Network (TCN) [38], which models temporal dependencies using one-dimensional convolutions with an expanding receptive field. TCNs have been widely adopted for video-based affect prediction [65, 69]. Finally, task-specific regression and classification heads produce predictions for the affective analysis tasks. An overview of the complete pipeline is shown in Fig. 3. The remainder of this section details the proposed attentive aggregation framework and its associated learning objectives.

3.1 Causal Supervision for Attention

To improve cross-subject robustness and further emphasize the most important regions for emotion recognition, we use the Causal Supervision for Attention (CSA) framework [61]. This framework introduces an explicit supervision signal motivated by causal inference. Given patch features X, attention weights A, and model predictions $\hat{Y}$, CSA models how attention influences the final prediction and measures the contribution of attention by comparing predictions before and after modifying the attention weights. Specifically, the factual prediction is represented as $\hat{Y}_{\mathbf{x},\mathbf{a}} = \hat{Y}(X = \mathbf{x}, A = \mathbf{a})$, while the counterfactual prediction is obtained by manually replacing the attention weights with counterfactual attention weights, $\hat{Y}_{\mathbf{x},\tilde{\mathbf{a}}} = \hat{Y}(X = \mathbf{x}, \mathrm{do}(A = \tilde{\mathbf{a}}))$, where the $\mathrm{do}(\cdot)$ operator [47] denotes an intervention that forces the attention to take the value $\tilde{\mathbf{a}}$ rather than the value it would naturally produce. This allows us to measure how the prediction changes when only the attention mechanism is altered (see Fig. 2). The

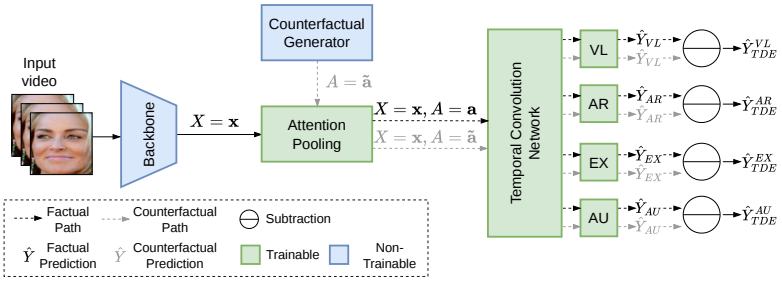

Fig. 3: Overview of the multi-task affect estimation pipeline. Feature representations are first extracted using a frozen backbone. Factual samples (dashed black arrows) and counterfactual samples (dashed grey arrows) are then generated and processed by a temporal encoder, followed by task-specific prediction heads for affect estimation. Finally Total Direct Effect is calculated per task.

difference between the learned attention weights and the counterfactual attention weights is measured using the Total Direct Effect (TDE) [59] (see Fig. 3), $\hat{Y}_{TDE} = \hat{Y}_{\mathbf{x},\mathbf{a}} - \hat{Y}_{\mathbf{x},\tilde{\mathbf{a}}}$, which captures the direct causal contribution of attention to the model output. CSA maximizes this effect during training by introducing the following auxiliary supervision objective:

$$\mathcal{L}_{CSA} = \mathcal{L}(\hat{Y}_{TDE}, y) \tag{1}$$

where y is the ground-truth label, and $\mathcal{L}$ is a generic loss function, as this method does not depend on any specific loss formulation.

Counterfactual generation. To generate counterfactual attention vectors, we sample a random vector from a standard Gaussian distribution and normalize it with a softmax operation so that its elements form a valid attention distribution, i.e., they sum to one [61]. Formally,

$$\tilde{\mathbf{a}} = \text{Softmax}(\mathbf{z}), \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

The resulting counterfactual attention may range from poor (e.g., focusing on background regions) to informative (e.g., emphasizing discriminative facial regions) [61]. By maximizing internal sensitivity over diverse counterfactual attention maps, the model is encouraged to consistently rely on discriminative facial regions, leading to more robust attention.

3.2 Cross-Covariance Regularization

To encourage complementary Key and Value representations in scaled dot-product attention, we introduce a cross-covariance regularization loss (Fig. 2) inspired by [4, 67]. Given the Key and Value representations, K and V, we first center them along the token dimension,

$$\mathbf{K}_c = \mathbf{K} - \frac{1}{S}\sum_{s=1}^{S}\mathbf{K}_s, \qquad \mathbf{V}_c = \mathbf{V} - \frac{1}{S}\sum_{s=1}^{S}\mathbf{V}_s, \tag{3}$$

where S is the number of tokens (or spatial locations). The cross-covariance $\mathbf{C}_{KV}$ and corresponding regularization loss $\mathcal{L}_{\text{cov}}$ are then:

$$\mathbf{C}_{KV} = \frac{1}{S}\mathbf{K}_c^{\top}\mathbf{V}_c, \qquad \mathcal{L}_{\text{cov}} = \|\mathbf{C}_{KV}\|_F^2 . \tag{4}$$

Minimizing $\mathcal{L}_{\text{cov}}$ reduces redundancy between the Key and Value feature spaces, encouraging complementary representations.

3.3 V projection

We introduce an architectural change by replacing the standard fully connected (linear) layer with a SwiGLU layer [55], a gated feed-forward unit that improves representational capacity by modulating one linear projection with another through a smooth non-linear gate. Formally,

$$\text{SwiGLU}(\mathbf{x}) = \big(\text{SiLU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\big) \odot (\mathbf{x}\mathbf{W}_2 + \mathbf{b}_2), \tag{5}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are learnable bias vectors, $\odot$ denotes element-wise multiplication, and $\text{SiLU}(\mathbf{x}) = \mathbf{x}\,\sigma(\mathbf{x})$ is the Sigmoid Linear Unit activation.

3.4 Training Objectives

We formulate the multi-task objective using task-specific losses. For valence-arousal estimation, we optimize the concordance correlation coefficient (CCC) based loss [31] together with mean squared error:

$$\begin{aligned}\mathcal{L}_{VA} = \mathcal{L}_{CCC}(\hat{y}_{VL}, y_{VL}) + \mathcal{L}_{MSE}(\hat{y}_{VL}, y_{VL}) \\ + \mathcal{L}_{CCC}(\hat{y}_{AR}, y_{AR}) + \mathcal{L}_{MSE}(\hat{y}_{AR}, y_{AR}),\end{aligned} \tag{6}$$

where y_{VL} and y_{AR} denote the ground-truth valence (VL) and arousal (AR) labels, and $\hat{y}_{VL}$ and $\hat{y}_{AR}$ denote their corresponding predictions. For expression recognition, we use cross entropy (CE):

$$\mathcal{L}_{EX} = CE(\hat{y}_{EX}, y_{EX}), \tag{7}$$

where y_{EX} and $\hat{y}_{EX}$ denote the ground-truth and predicted expression labels. For AU detection, we employ binary cross entropy (BCE):

$$\mathcal{L}_{AU} = BCE(\hat{y}_{AU}, y_{AU}), \tag{8}$$

where y_{AU} and $\hat{y}_{AU}$ represent the ground-truth and predicted AU labels. The joint task objective is defined as:

$$\mathcal{L}_{Task} = \mathcal{L}_{AU} + \mathcal{L}_{EX} + \mathcal{L}_{VA}. \tag{9}$$

Additionally, we incorporate auxiliary supervision through the CSA loss:

$$\mathcal{L}_{CSA} = \mathcal{L}_{CSA}^{AU} + \mathcal{L}_{CSA}^{EX} + \mathcal{L}_{CSA}^{VA}, \tag{10}$$

where each term corresponds to the CSA objective in Eq. 1 of the respective task. The final optimization objective is:

$$\mathcal{L}_{Overall} = \mathcal{L}_{Task} + \lambda_{CSA}\mathcal{L}_{CSA} + \lambda_{cov}\mathcal{L}_{cov}. \tag{11}$$

where the coefficients λ_{CSA} and λ_{cov} scale the contributions of the CSA (Eq. 1) and cross-covariance regularization (Eq. 4) terms.

4 Experiments

To evaluate the effectiveness of our proposed framework in learning subject-independent affect representations, we compare it against standard attentive pooling and conduct experiments with two backbone networks pre-trained on different datasets, further demonstrating its robustness. The first backbone is FRoundation [9], specialised for this challenge and pre-trained on facial images. The second backbone is DINOv2 [45], pre-trained on a general-purpose image corpus. We compare the performance of both backbones against the official baseline, which uses a pre-trained ConvNeXt [41] model with frozen convolutional weights and MixAugment [49] data augmentation, on the official validation set. To further assess the generalisability of our method, we perform additional experiments using 5-fold cross-validation. Finally, to investigate the contribution of each component within our framework, we conduct a comprehensive ablation study. The remainder of this section describes the dataset used in this study, the evaluation metrics, and the implementation details.

4.1 Dataset

For the MTL track, the organizers provide 142,382 training images and 26,876 validation images from the Aff-Wild2 dataset [31]. Continuous VA annotations are provided in the range $[-1, 1]$. EX annotations consist of eight categories: Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Other. AU annotations comprise 12 classes: AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, and AU26. Following the removal of samples with invalid annotations (AU labels of -1, VA labels of -5, or EX labels of -1), the resulting training and validation sets contain 52,154 and 15,440 images, respectively.

4.2 Metrics

The performance measure P as specified by the ABAW MTL challenge is calculated by summing the following components: the mean CCC for valence and arousal, the average F_1 Score across all eight expression categories F_{EX}, and the average F_1 Score across all 12 action units F_{AU}. It is mathematically formulated as follows:

$$P = \frac{CCC_{AR} + CCC_{VL}}{2} + F_{EX} + F_{AU}. \tag{12}$$

4.3 Implementation Details

Our method is implemented in PyTorch [46], all experiments are performed on a single NVIDIA RTX A6000 GPU using cropped and aligned facial images resized to 224×224 pixels [31]. For feature extraction, we utilize two models sharing a ViT-S architecture (16×16 patches, 256 tokens, 384 embedding dimensions): FRoundation [9], trained on WebFace4M with the ArcFace loss [11, 70], and DINOv2 [45], pre-trained on the general LVD-142M dataset. The TCN module consists of 3 layers with a kernel size of 5. The attention head size is 64, while the number of heads matches the backbone output dimension. Training uses the AdamW optimizer [42] with a batch size of 8, an initial learning rate of 8×10^{-5}, linear warmup for the first 10% of steps, and a cosine annealing schedule. Models are trained for 15 epochs with early stopping (patience of 5) monitored on a random validation split comprising 15% of the training data.. The coefficients λ_{CSA} and λ_{cov} in Eq. 11 were selected via preliminary grid search over $[0.01, 1.0]$, using finer steps at lower values and coarser steps at higher values. The final values were $\lambda_{CSA} = 0.1$ $\lambda_{cov} = 0.05$ for DINOv2 and $\lambda_{CSA} = 0.35$ $\lambda_{cov} = 0.15$ for FRoundation.

5 Results

Comparison with the official baseline and five-fold cross-validation. Table 1 presents a comparison with the official baseline and the results of five-fold cross-validation, where Fold 1 corresponds to the official validation split. The ConvNeXt baseline achieves an overall $P = 0.45$ (row 11), while our method substantially improves upon this result, achieving $P = 0.96$ (row 1) with the DINOv2 backbone and $P = 1.22$ (row 6) with the FRoundation backbone. Across all cross-validation folds, FRoundation consistently outperforms DINOv2. The most substantial gains are observed in VA estimation, where FRoundation increases CCC_{VA} by up to 0.225 (Fold 5). Expression recognition demonstrates a similarly consistent improvement, yielding higher F_{EX} scores across all folds, with gains up to 0.106 on Fold 3. Conversely, AU detection remains comparable between the models, exhibiting only minor variations.

Ablation study. Table 2 presents an ablation study of the proposed framework. Results are reported as the mean and 95% confidence interval of P using

Table 1: The results of 5-fold cross-validation, Fold 1 corresponds to the official validation set. Final row corresponds to the official baseline on the validation set.

Backbone	Fold	CCC_{VL}	CCC_{AR}	CCC_{VA}	F_{EX}	F_{AU}	P
DINOv2	1	0.3683	0.2663	0.3173	0.2504	0.3982	0.9658
	2	0.3726	0.3466	0.3596	0.2497	0.3849	0.9942
	3	0.2029	0.2906	0.2468	0.1919	0.3688	0.8074
	4	0.1748	0.4352	0.3050	0.2162	0.3696	0.8907
	5	0.2712	0.5597	0.4155	0.2816	0.4294	1.1264
FRoundation	1	0.5299	0.4947	0.5123	0.3116	0.3974	1.2214
	2	0.4726	0.3058	0.3892	0.3044	0.3758	1.0694
	3	0.5565	0.3952	0.4758	0.2974	0.3989	1.1722
	4	0.5165	0.3697	0.4431	0.3129	0.3694	1.1254
	5	0.6008	0.6797	0.6403	0.3063	0.4392	1.3858
ConvNeXt	1	-	-	-	-	-	0.4500

Table 2: Ablation study of the proposed framework. Results are reported as mean ± 95% CI across 5-fold cross-validation. The highest mean performance is highlighted in bold. SwiGLU indicates the presence of the non-linear projection, CSA denotes the causal supervision applied to the attention mechanism, and Cov denotes the covariance regularization term. Checkmarks indicate that the corresponding component is enabled in the configuration. For conciseness, only the overall performance metric P is reported.

Backbone	SwiGLU	CSA	Cov	P
DINOv2	×	×	×	0.8730±0.0912
	×	✓	×	0.8862±0.1363
	×	×	✓	0.8686±0.1049
	×	✓	✓	**0.9106±0.1260**
	✓	×	×	0.8616±0.1207
	✓	✓	×	0.9036±0.1037
	✓	×	✓	0.9317±0.2013
	✓	✓	✓	**0.9569±0.1481**
FRoundation	×	×	×	1.0524±0.0861
	×	✓	×	1.1043±0.0610
	×	×	✓	1.0784±0.1431
	×	✓	✓	**1.1212±0.0987**
	✓	×	×	1.1265±0.1583
	×	✓	×	1.1555±0.1004
	×	×	✓	1.1588±0.1903
	×	✓	✓	**1.1948±0.1498**

5-fold cross-validation with the identical splits from prior experiments. The baseline configuration consists of a linear Value projection without causal supervision of attention or covariance regularization (DINOv2 first row, FRoundation row 9). Adding either causal supervision or covariance regularization generally improves performance. The only exception is the DINOv2 backbone with a linear Value projection, where covariance regularization causes a slight performance drop.

Table 3: Results on the official test set, and comparison with other contestants.

Team	CCC_{VA}	F_{EX}	F_{AU}	P
baseline	0.1473	0.1517	0.1710	0.4700
CPR	0.5347	0.3625	0.5532	1.4504
HackFleet	0.5142	0.3415	0.5337	1.3893
EagleonPamir1 [57]	0.4966	0.3479	0.5150	1.3596
HSEmotion [2]	0.4068	0.3444	0.5232	1.2744
CVPD [5]	0.3141	0.2835	0.4858	1.0833
ours	0.3482	0.3189	0.4245	1.0916

A possible explanation is that, with the weaker DINOv2 representations and a linear Value projection, regularization overly limits learned features. Combining both components consistently improves performance regardless of the Value transformation. Replacing the linear Value projection with SwiGLU further improves performance for FRoundation (row 9 vs row 13) but slightly degrades performance for DINOv2 (row 1 vs row 5), again suggesting that the weaker DINOv2 representations are unable to fully exploit the additional capacity. However, combining SwiGLU with causal supervision and covariance regularization improves performance for both backbones. This suggests that covariance regularization is more effective when paired with the more expressive SwiGLU transformation, which provides greater flexibility to learn decorrelated representations while preserving task-relevant information. The complete framework achieves the best overall performance, improving over conventional attention pooling by 0.08 (row 1 vs row 8) and 0.142 (row 9 vs row 16) for DINOv2 and FRoundation, respectively. Overall, the results show that each proposed component contributes to performance, while their combination yields the largest and most consistent gain with both backbone architectures.

Evaluation on the test set. Table 3 reports the results obtained on the test set of the MTL challenge in the 11th ABAW Competition, together with a comparison against the participating teams. The reported results for our method are based on the FRoundation backbone. Our approach substantially outperforms the official baseline, achieving a composite P score that is more than twice as high (1.09 compared to 0.47), and ranks fifth among the six participating teams. However, this ranking should be interpreted in light of the methodological differences across submissions. The teams ranked first and second employ model ensembling, while the teams ranked third and fourth combine multiple backbones through feature fusion and further apply post-hoc refinements such as prediction smoothing. In contrast, only CVPD (the sixth-place team) adopts a setting comparable to ours, namely a single-backbone architecture without additional post-hoc adaptation. Under this comparable setting, our method achieves slightly better performance compared to CVPD on valence-arousal estimation and expression recognition, and also attains a higher overall composite score P.

6 Discussion

The proposed approach is influenced by the representational capacity and potential inherent biases of the pre-trained backbones used, which may impact its performance and generalisation across different settings. Additionally, the counterfactual generation process explored in this work is limited to a single strategy that does not incorporate the current values of the attention weights. The K, V regulariser is based on cross-covariance, allowing it to capture linear relationships within the data, but limiting its ability to model more complex non-linear dependencies. While the introduction of the SwiGLU layer provides a mechanism to explore these non-linear relationships, it also increases computational overhead due to its approximately doubled parameter count compared to a standard fully connected layer.

Future work offers several directions to extend both our evaluation and core methodology. In particular, attention consistency and resilience to spurious visual factors could be more rigorously quantified by systematically altering illumination, applying pose rotations, synthetically editing frames to introduce facial accessories, and creating additional hand-labeled demographic annotations. This study did not investigate the use of task-specific learnable query vectors, as explored in [39], where each query vector receives causal supervision tailored to its corresponding task. Such an approach could provide a more targeted mechanism for extracting task-relevant information. The core methodology could also be improved by replacing the current counterfactual generation strategy with informed search procedures over the simplex defined by the attention weights, enabling more structured exploration of alternative attention distributions. Furthermore, non-linear dependencies between the K and V representations could be identified and regularised using kernel-based measures such as the Hilbert-Schmidt Independence Criterion [15, 44]. The generalisability of the approach could be further assessed on additional datasets, such as AFEW-VA [37] and RECOLA [50].

7 Conclusion

In this work, we integrated causally inspired learning into vision-based affective behaviour analysis to improve the robustness of facial expression understanding in real-world settings, addressing the MTL challenge of the 11th ABAW Competition. We further introduced an independence regularization mechanism to encourage more diverse representations between the K and V attention projections, and replaced the conventional linear V projection with a SwiGLU layer to increase representational capacity and better capture the complementary information promoted by the regularization. Quantitative results and attention visualizations (Fig. 1) demonstrate that our approach directs attention toward affect-relevant facial regions while suppressing spurious factors such as illumination, pose, and demographic variations. Overall, our method achieved 5th place in the MTL challenge of the 11th ABAW competition, outperforming the only

other non-ensemble approach, and improved upon standard attention pooling by up to 0.142 on the composite metric P. Extensive ablation studies demonstrate the contribution of each proposed component, highlighting the necessity of their inclusion.

Acknowledgements

This project has received funding from the European Union's Horizon Europe programme (grant agreement No. 101178988). Nemanja Rašajski and Konstantinos Makantasis were supported by the GenV project (REP-2025-032) funded by Xjenza Malta.

References

1. Ayata, D., Yaslan, Y., Kamasak, M.E.: Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. Journal of Medical and Biological Engineering **40**(2), 149–157 (2020)
2. Bakin, A., Savchenko, A.V.: HSEmotion team at the 11th ABAW challenge: Multi-task learning and ambivalence/hesitancy video recognition. arXiv preprint arXiv:2607.12774 (2026)
3. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. Transactions on Machine Learning Research (2024)
4. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-invariance-covariance regularization for self-supervised learning. In: Proceedings of the International Conference on Learning Representations (2022)
5. Bekhouche, S.E., Sellam, A.Z., Dornaika, F., Hadid, A.: AffectFlow-DINO: Uncertainty-aware multi-task affect estimation via conditional rectified flow. arXiv preprint arXiv:2607.13250 (2026)
6. Cabacas-Maso, J., Ortega-Beltrán, E., Benito-Altamirano, I., Ventura, C.: Enhancing facial expression recognition through dual-direction attention mixed feature networks: Application to 7th ABAW challenge. In: Proceedings of the European Conference on Computer Vision. pp. 311–321. Springer (2024)
7. Canal, F.Z., Müller, T.R., Matias, J.C., Scotton, G.G., de Sa Junior, A.R., Pozzebon, E., Sobieranski, A.C.: A survey on facial emotion recognition techniques: A state-of-the-art literature review. Information Sciences **582**, 593–617 (2022)
8. Chen, Y., Yang, X., Cham, T.J., Cai, J.: Towards unbiased visual emotion recognition via causal intervention. In: Proceedings of the ACM International Conference on Multimedia. pp. 60–69 (2022)
9. Chettaoui, T., Damer, N., Boutros, F.: FRoundation: Are foundation models ready for face recognition? Image and Vision Computing **156** (2025)
10. Deng, D., Chen, Z., Shi, B.E.: Multitask emotion recognition with incomplete labels. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. pp. 592–599. IEEE (2020)
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)

12. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978)
13. Ekman, P., Friesen, W.V., et al.: The Repertoire of Nonverbal Behavior: Categories, Rrigins, Usage and Coding, vol. 1. Mouton de Gruyter Berlin (1969)
14. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
15. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: Proceedings of the International Conference on Algorithmic Learning Theory. pp. 63–77. Springer (2005)
16. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions **1**(1), 68–99 (2010)
17. Guo, R., Guo, H., Wang, L., Chen, M., Yang, D., Li, B.: Development and application of emotion recognition technology—a systematic literature review. BMC psychology **12**(1) (2024)
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Proceedings of the NeurIPS Deep Learning Workshop (2014)
19. Kollias, D., Schulc, A., Hajiyev, E., Zafeiriou, S.: Analysing affective behavior in the first ABAW 2020 competition. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. pp. 794–800 (2020)
20. Kollias, D.: ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2328–2336 (2022)
21. Kollias, D.: ABAW: Learning from synthetic data & multi-task learning challenges. In: Proceedings of the European Conference on Computer Vision. pp. 157–172. Springer (2023)
22. Kollias, D., Shao, C., Kaloidas, O., Patras, I.: Behaviour4all: in-the-wild facial behaviour analysis toolkit. arXiv preprint arXiv:2409.17717 (2024)
23. Kollias, D., Sharmanska, V., Zafeiriou, S.: Face behavior a la carte: Expressions, affect and action units in a single network. arXiv preprint arXiv:1910.11111 (2019)
24. Kollias, D., Sharmanska, V., Zafeiriou, S.: Distribution matching for heterogeneous multi-task learning: a large-scale face study. arXiv preprint arXiv:2105.03790 (2021)
25. Kollias, D., Sharmanska, V., Zafeiriou, S.: Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2813–2821 (2024)
26. Kollias, D., Tzirakis, P., Baird, A., Cowen, A., Zafeiriou, S.: ABAW: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5888–5897 (2023)
27. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Baird, A., Gagne, C., Shao, C., Hu, G.: The 6th affective behavior analysis in-the-wild (ABAW) competition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4587–4598 (2024)
28. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Granger, E., Pedersoli, M., Bacon, S., Baird, A., Gagne, C., et al.: Advancements in affective and behavior analysis: The 8th ABAW workshop and competition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5572–5583 (2025)
29. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Granger, E., Pedersoli, M., Bacon, S., Madsen, J., Belharbi, S., et al.: From affect to complex behavior:

Advancing multimodal human-centered ai at the 10th ABAW workshop & competition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5302–5311 (2026)
30. Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision pp. 1–23 (2019)
31. Kollias, D., Zafeiriou, S.: Aff-Wild2: Extending the Aff-Wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)
32. Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855 (2019)
33. Kollias, D., Zafeiriou, S.: Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. arXiv preprint arXiv:2103.15792 (2021)
34. Kollias, D., Zafeiriou, S.: Analysing affective behavior in the second ABAW2 competition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3652–3660 (2021)
35. Kollias, D., Zafeiriou, S., Kotsia, I., Dhall, A., Ghosh, S., Shao, C., Hu, G.: 7th ABAW competition: Multi-task learning and compound expression recognition. In: Proceedings of the European Conference on Computer Vision. pp. 31–45. Springer (2024)
36. Kollias, D., Zafeiriou, S., Kotsia, I., Slabaugh, G., Senadeera, D.C., Zheng, J., Yadav, K.K.K., Shao, C., Hu, G.: From emotions to violence: Multimodal fine-grained behavior analysis at the 9th ABAW. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–12 (2025)
37. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M.: AFEW-VA database for valence and arousal estimation in-the-wild. Image and Vision Computing **65**, 23–36 (2017)
38. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
39. Li, X., Du, W., Yang, H.: Affective behavior analysis using task-adaptive and au-assisted graph. In: Proceedings of the European Conference on Computer Vision. pp. 393–403. Springer (2024)
40. Liu, C., Zhang, W., Qiu, F., Li, L., Wang, D., Yu, X.: Affective behaviour analysis via progressive learning. In: Proceedings of the European Conference on Computer Vision. pp. 366–379. Springer (2024)
41. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
42. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
43. Marinoiu, E., Zanfir, M., Olaru, V., Sminchisescu, C.: 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2158–2167 (2018)
44. Mooij, J., Janzing, D., Peters, J., Schölkopf, B.: Regression by dependence minimization and its application to causal inference in additive noise models. In: Proceedings of International Conference on Machine Learning. pp. 745–752 (2009)
45. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

46. Paszke, A., et al.: Automatic differentiation in Pytorch. In: Proceedings of the NeurIPS Workshop Autodiff (2017)
47. Pearl, J.: Causality. Cambridge university press (2009)
48. Picard, R.W.: Affective computing. MIT press (2000)
49. Psaroudakis, A., Kollias, D.: Mixaugment & mixup: Augmentation methods for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2367–2375 (2022)
50. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. pp. 1–8. IEEE (2013)
51. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6) (1980)
52. Sajjad, M., Ullah, F.U.M., Ullah, M., Christodoulou, G., Cheikh, F.A., Hijji, M., Muhammad, K., Rodrigues, J.J.: A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. Alexandria Engineering Journal **68**, 817–840 (2023)
53. Savchenko, A.V.: HSEmotion team at the 7th ABAW challenge: Multi-task learning and compound facial expression recognition. arXiv preprint arXiv:2407.13184 (2024)
54. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
55. Shazeer, N.: GLU variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
56. Spezialetti, M., Placidi, G., Rossi, S.: Emotion recognition for human-robot interaction: Recent advances and future perspectives. Frontiers in Robotics and AI **7** (2020)
57. Sun, J., Gao, Z.: Task-specific feature fusion method for multi-task affective behavior analysis. arXiv preprint arXiv:2607.13986 (2026)
58. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nature Machine Intelligence **3**(1), 42–50 (2021)
59. VanderWeele, T.J.: Explanation in causal inference: developments in mediation and interaction. International journal of epidemiology **45**(6), 1904–1908 (2016)
60. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proceedings of the Advances in neural information processing systems **30** (2017)
61. Wang, H., Chen, J., Du, L., Fu, Q., Han, S., Song, X.: Causal-based supervision of attention in graph neural network: a better and simpler choice towards powerful attention. In: Proceedings of the International Joint Conference on Artificial Intelligence (2023)
62. Wang, T., Zhou, C., Sun, Q., Zhang, H.: Causal attention for unbiased visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3091–3100 (2021)
63. Xia, R., Liu, Y.: A multi-task learning framework for emotion recognition using 2d continuous space. IEEE Transactions on Affective Computing **8**(1), 3–14 (2015)
64. Yang, D., Chen, Z., Wang, Y., Wang, S., Li, M., Liu, S., Zhao, X., Huang, S., Dong, Z., Zhai, P., et al.: Context de-confounded emotion recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19005–19015 (2023)

65. Yu, J., Wang, Y., Wang, L., Zheng, Y., Xu, S.: Interactive multimodal fusion with temporal modeling. arXiv preprint arXiv:2503.10523 (2025)
66. Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-Wild: Valence and arousal 'in-the-wild'challenge. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 1980–1987. IEEE (2017)
67. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Proceedings of the International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
68. Zhang, S., Zhang, Y., Zhang, Y., Wang, Y., Song, Z.: A dual-direction attention mixed feature network for facial expression recognition. Electronics **12**(17), 3595 (2023)
69. Zhou, W., Lu, J., Xiong, Z., Wang, W.: Leveraging tcn and transformer for effective visual-audio fusion in continuous emotion recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5756–5763 (2023)
70. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., et al.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10492–10502 (2021)