Behaviour4All: A Dependency-Aware Toolkit for in-the-wild Facial Behaviour Analysis

Dimitrios Kollias[1,2*] **Chunchang Shao**[1] **Odysseus Kaloidas** **Ioannis Patras**[1]
[1]Center for Multimodal AI, Queen Mary University of London, UK
[2]Digital Environment Research Institute, Queen Mary University of London, UK
d.kollias@qmul.ac.uk

Abstract

Facial behaviour analysis requires understanding valence–arousal (VA), expressions (EXP), action units (AUs), yet existing systems treat these tasks independently or use naïve MTL, ignoring their structured dependencies. This leads to negative transfer, conflicting gradients, poor cross-dataset generalisation, especially with heterogeneous datasets with only partial labels. We propose Behaviour4All, the first dependency-aware toolkit that unifies VA, EXP, AUs via explicit relational priors and principled optimisation framework. Behaviour4All integrates psychophysically grounded task mappings with coordinated loss families to mitigate gradient conflicts and enforce coherent cross-task semantics. Across 13 in-the-wild datasets, Behaviour4All achieves state-of-the-art performance, improves fairness and generalisation, and outperforms vanilla MTL and student–teacher baselines. The toolkit is publicly available at: https://github.com/dkollias/Behaviour4All.

1 Introduction

Understanding human facial behaviour requires modelling several complementary affective representations. Action Unit Detection (AUD) describes local facial muscle activations based on the Facial Action Coding System [10]; Basic Expression Recognition (BER) assigns faces to discrete emotional categories such as happiness, sadness or anger; and Valence–Arousal Estimation (VA-E) represents affect in a continuous space of emotional polarity and intensity [50]. These tasks capture different views of the same underlying affective state: AUs provide fine-grained muscular evidence, expressions provide categorical semantic interpretation, and VA provides a continuous affective description. A robust facial behaviour analysis system should therefore not treat these tasks as independent, but should model their structured dependencies. Despite substantial progress in automatic behaviour analysis (ABA), existing methods and toolkits remain limited in several important ways. First, many systems are trained and evaluated under narrow task settings, often focusing on only one or two affective outputs. Second, most datasets provide annotations for only a subset of tasks, resulting in heterogeneous and partially overlapping supervision. Third, standard multitask learning (MTL) typically optimises a shared backbone using a weighted sum of task losses, without explicitly modelling task dependencies. This makes training vulnerable to negative transfer [55], conflicting gradients and semantic drift between task heads, especially when VA, expressions and AUs are trained from different datasets with non-overlapping annotations. Finally, existing toolkits often lack support for VA estimation, rely on separate models for different tasks, or use traditional machine learning components rather than a unified deep learning framework.

In this work, we introduce **Behaviour4All**, a dependency-aware MTL framework and toolkit for ABA. Rather than treating VA, EXP and AUs as separate or weakly related outputs, Behaviour4All reformulates MT facial behaviour analysis as a structured optimisation problem that explicitly

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

exploits dependencies across continuous, categorical and multi-label representations. The framework integrates three complementary components. Task-specific supervision preserves the discriminative performance of individual tasks. Cross-task dependency modelling leverages relational priors to establish bidirectional consistency between task predictions, enabling structured supervision transfer under heterogeneous and partially overlapping annotations, with confidence-adaptive regulation of cross-task constraints. Geometry modelling aligns semantically compatible representations and separates incompatible ones in the affective and shared feature spaces. Crucially, these components are jointly optimised to enforce coherence at both the prediction and representation levels, mitigating semantic drift, negative transfer and conflicting gradients. Overall, our contributions are as follows:

- *A Dependency-Aware MTL Framework for Facial Behaviour Analysis:* We reformulate facial behaviour analysis as a structured multitask optimisation problem grounded in task-specific supervision, cross-task dependency modelling and geometry modelling.
- *Explicit Relational and Geometric Modelling of VA, EXP, AUs:* We introduce relational priors and geometry-aware losses that encode dependencies between continuous VA, categorical EXP, binary AU spaces, reducing negative transfer and conflicting gradients under heterogeneous supervision.
- *A Complete Toolkit with Comprehensive Evaluation:* We instantiate the framework as Behaviour4All, an open-source, real-time toolkit supporting face detection, alignment, VA estimation, EXP recognition, AU detection. We evaluate it on thirteen in-the-wild datasets in terms of overall performance, fairness, cross-dataset generalisation, downstream utility and efficiency. Behaviour4All consistently outperforms all sota and existing toolkits in all these evaluations in all datasets and tasks. Compared to existing toolkits, Behaviour4All achieves a greater speedup, delivering superior efficiency without compromising performance.

Table 1: Comparison of facial behaviour analysis toolkits

Toolkit	Bbox	Landmarks	AU	BER	VA	Free	GPU	Train	Test	MTL	wild dbs	multiple dbs	Downstream
AFFDEX 2.0 [3]	✓	✓	✓	✓							✓		
FACET	✓	✓	✓	✓									
OpenFace 2.0 [1]	✓	✓	✓			✓		✓	✓			✓	
LibreFace [5]			✓	✓		✓	✓	✓	✓		✓		
py-feat [7]	✓	✓	✓	✓		✓	✓		✓		✓	✓	
Behaviour4All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2 Related Work

Toolkits Table 1 presents an overview of recently developed ABA toolkits. *Bbox* specifies whether the toolkit includes built-in face detection or relies on external face detection software. *Landmarks* indicates whether landmark localization is performed, which is crucial for accurate face alignment. We note that LibreFace depends critically on (the external) Mediapipe, which introduces: (i) stability risks (Mediapipe components are frequently updated, deprecated or replaced); (ii) version inconsistency (different installations yield different detections and downstream predictions); (iii) limited controllability. *AU/BER/VA* denotes whether the toolkit supports AUD/BER/VA-E. *Free* specifies whether the toolkit is freely available. *GPU* specifies whether the toolkit provides support for GPU. *Train/Test* indicates the availability of training/re-training code and whether pretrained models and inference scripts are provided. *MTL* identifies whether the toolkit employs a single multi-task model for all behaviour analysis tasks. Behaviour4All is the only toolkit that simultaneously handles AUD, BER, VA-E within a unified model; all other toolkits rely on separate models for each task and address at most two tasks. *wild dbs* indicates whether the toolkit has been trained on in-the-wild datasets. Behaviour4All is the only toolkit that utilizes in-the-wild datasets for all tasks, including AUD. *multiple dbs* specifies whether the toolkit has been trained on multiple datasets. *Downstream* evaluates whether the toolkit has been tested on downstream tasks or if its generalizability has been validated on external datasets. Behaviour4All is the only toolkit that fulfills this criterion.

SOTA In appendix, we provide detailed information of various uni-/multi-task sota for BER, AUD, VA-E, e.g., DAN [56], MA-Net [69], EAC [67], VA-StarGan [28], MT-EffNet [51], POSTER++ [41], UA-FER [70], MHAN [53], CLIPER [36], CLIP-FT [47], FMAE-IAT[46], AUFormer [61], MDHR [54], as well as FUXI [65], SITU [38], CTC [72], USTC-IAT[60] GRJCA[48] CtyunAI[71]. Existing methods focus on one-directional or auxiliary relationships, do not generalise across tasks and do not address MTL under heterogeneous supervision.

3 Behaviour4All Toolkit

Behaviour4All (can be seen in Fig. 1) has 2 components: FaceLocalizationNet & FaceMultiBehaviourNet. The first one performs image pre-processing, which involves simultaneous Face and Landmark Detection and Image Alignment. The pre-processed images are then fed to FaceMultiBehaviourNet that performs simultaneous Detection of 17 AUs, Recognition of 7 Basic Expressions and Estimation of Valence-Arousal. These components are explained next.

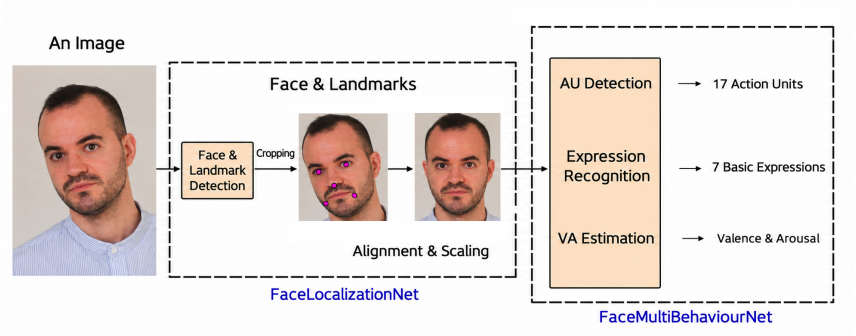

Figure 1: Behaviour4All Toolkit, composed of: FaceLocalizationNet (simultaneous Face and Landmark Detection & Alignment) and FaceMultiBehaviourNet (simultaneous 17 AUs Detection, 7 Basic Expression Recognition & Valence-Arousal Estimation)

FaceLocalizationNet For face detection and alignment, we utilize RetinaFace [8], a sota single-stage face detector that directly predicts face locations along with 5 key landmarks (eyes, nose tip, mouth corners). Its design, incorporating a feature pyramid network, context head module, cascade MT loss, ensures high accuracy even under challenging conditions (variations in lighting, pose, occlusion, scale). RetinaFace extracts multi-scale features with CNN backbone. In our toolkit, we offer two backbone options: MobileNet-0.25 for lightweight applications and ResNet-50 for higher accuracy. Facial alignment is performed using the 5 localized landmarks. This process applies geometric transformations (translation, rotation, scaling) to standardize the face, ensuring consistent feature positioning. The aligned faces are resized to $112 \times 112 \times 3$.

FaceMultiBehaviourNet is a Multi-Task (MT) CNN that simultaneously performs: (i) estimation of Valence and Arousal (VA); (ii) recognition of 7 basic facial expressions; and (iii) detection of 17 binary facial Action Units (AUs). For an image, available annotations may include: basic expressions $y_{\text{exp}} \in \{1, \ldots, 7\}$; AUs $y_{\text{AU}} \in \{0, 1\}^{17}$; valence ($y_{\text{V}} \in [-1, 1]$) and arousal ($y_{\text{A}} \in [-1, 1]$).

Dependency-aware multitask optimisation. Behaviour4All differs from standard MTL by reformulating facial behaviour analysis as a structured dependency-aware optimisation problem. Conventional MTL typically relies on a shared backbone and optimises a weighted sum of task-specific losses, without explicitly modelling the semantic and geometric dependencies between tasks. Such formulations often assume fully annotated datasets (which can be impractical), require manually tuned or heuristic loss balancing strategies (which is ad hoc, manual and computationally expensive), and remain vulnerable to negative transfer, conflicting gradients and poor generalisation when tasks are trained from heterogeneous datasets with partially overlapping annotations. Although several MTL variants address these issues through uncertainty weighting, student–teacher schemes or more complex task-specific architectures, they often introduce additional computational cost and do not explicitly encode the relational structure between VA, EXP and AUs.

In contrast, Behaviour4All is built on three principled components. First, *task-specific supervision* preserves strong individual task performance for VA estimation, EXP recognition and AU detection. Second, *cross-task dependency modelling* explicitly encodes structured relations between affective tasks through relational priors, enabling learning under heterogeneous and partially overlapping annotations while enforcing semantic consistency across tasks. Third, *geometry modelling* structures the shared representation space so that compatible affective concepts are aligned and incompatible ones are separated. Together, these components directly address key limitations of standard MTL: negative transfer, conflicting gradients, partial supervision and weak cross-dataset generalisation.

Tasks Relationships (1) *EXP-VA:* Task relationship is inferred from AffectNet that includes both VA and EXP labels. For each EXP, we compute the centroid c_{exp} as the mean of all VA values of training images labeled with that EXP:

$$c_{exp} = \left(\frac{1}{K} \sum_{i=1}^{K} y_V^i \, , \frac{1}{K} \sum_{i=1}^{K} y_A^i \right) \tag{1}$$

where: y_V^i/y_A^i are V/A labels of image with EXP label; K is total number of images with EXP label. (2) *EXP-AUs:* [9] establishes relationship between EXP and AUs, identifying characteristic AUs consistently activated in human expressive behavior. Table 2 summarizes these relationships, which we leverage to define a structured task-relationship model. We categorize AUs in: *prototypical*, which are consistently marked as active across all annotators, and *observational*, which are activated by only a subset of annotators. Although alternative studies report slightly different AU–EXP associations, our framework remains flexible. As shown in supplementary, substituting an alternative mapping yields negligible performance changes, demonstrating our method's robustness. For *Behaviour4All*, we adopt the AU–EXP relationships in Table 2.

Table 2: EXP-AU Relationship; in parenthesis are weights denoting fraction of annotators that scored the AU activation

Expression	Prototypical AUs	Observational AUs
happiness	12, 25	6 (0.51)
sadness	4, 15	1 (0.6), 6 (0.5), 11 (0.26), 17 (0.67)
fear	1, 4, 20, 25	2 (0.57), 5 (0.63), 26 (0.33)
anger	4, 7, 24	10 (0.26), 17 (0.52), 23 (0.29)
surprise	1, 2, 25, 26	5 (0.66)
disgust	9, 10, 17	4 (0.31), 24 (0.26)

Overall Objective Function The final training objective is organised according to the three components of our dependency-aware MTL framework: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Tasks}} + \mathcal{L}_{\text{Priors}} + \mathcal{L}_{\text{Geometry}}$. $\mathcal{L}_{\text{Tasks}}$ provides direct task-specific supervision and ensures that each task remains individually discriminative. $\mathcal{L}_{\text{Priors}}$ performs cross-task dependency modelling by encoding structured relations between VA, EXP, AUs, enforcing semantic and functional consistency across partially annotated datasets. $\mathcal{L}_{\text{Geometry}}$ shapes shared latent space by aligning compatible task prototypes and separating incompatible affective regions, reducing representational conflict, improving cross-db generalisation.

Architecture Fig. 2 illustrates the architecture of FaceMultiBehaviourNet and the proposed dependency-aware optimisation framework, highlighting the corresponding task-specific, relational-prior and geometry-based losses. Each of these components is described in detail below.

Tasks Loss $\mathcal{L}_{\text{Tasks}} = \lambda_{\text{VA}}\mathcal{L}_{\text{VA}} + \lambda_{\text{EXP}}\mathcal{L}_{\text{EXP}} + \lambda_{\text{AU}}\mathcal{L}_{\text{AU}}$, where: $\mathcal{L}_{\text{EXP}}$ is Cross Entropy (CE) for BER. $\mathcal{L}_{AU}$ is weighted BCE for AUD. $\mathcal{L}_{VA} = 1 - 0.5 \cdot (CCC_A + CCC_V)$ is Concordance Correlation Coefficient (CCC) for VA-E.

Priors Loss It consists of three losses: (i) VA-EXP; (ii) AU–EXP; and (iii) Confidence-Weighted Distribution Synchronization. All terms are based on the prior knowledge about tasks' relationships.

VA–EXP loss VA and EXP provide complementary descriptions of affect: VA represents emotional state in a continuous space, whereas EXP provides a discrete semantic abstraction. Although these representations are strongly related, they are often learned independently or only weakly coupled, leading to semantic inconsistencies, e.g., predicting "happy" with negative or near-neutral valence, and unstable optimisation under heterogeneous supervision. Our objective is to ensure that VA and EXP predictions describe the *same underlying affective state* from different perspectives. VA estimates should be semantically compatible with EXP predictions, EXP categories should occupy well-structured and separable regions of the VA space, and mappings between VA and EXP should remain mutually consistent. To this end, the VA–EXP loss enforces three forms of consistency: semantic, geometric and functional. These terms stabilise learning under partial supervision, reduce cross-task disagreement and encourage a coherent affect manifold in which continuous and categorical representations remain mutually consistent and interpretable:

$$\mathcal{L}_{\text{VA–EXP}} = \lambda_{\text{s}}\mathcal{L}_{\text{semantic}} + \lambda_{\text{g}}\mathcal{L}_{\text{VA-geometry}} + \lambda_{\text{f}}\mathcal{L}_{\text{functional}}. \tag{2}$$

$\mathcal{L}_{semantic}$*:* This term aligns VA predictions with EXP at semantic level. It discourages semantically incompatible predictions, such as assigning high probability to "happy" when V is strongly negative.

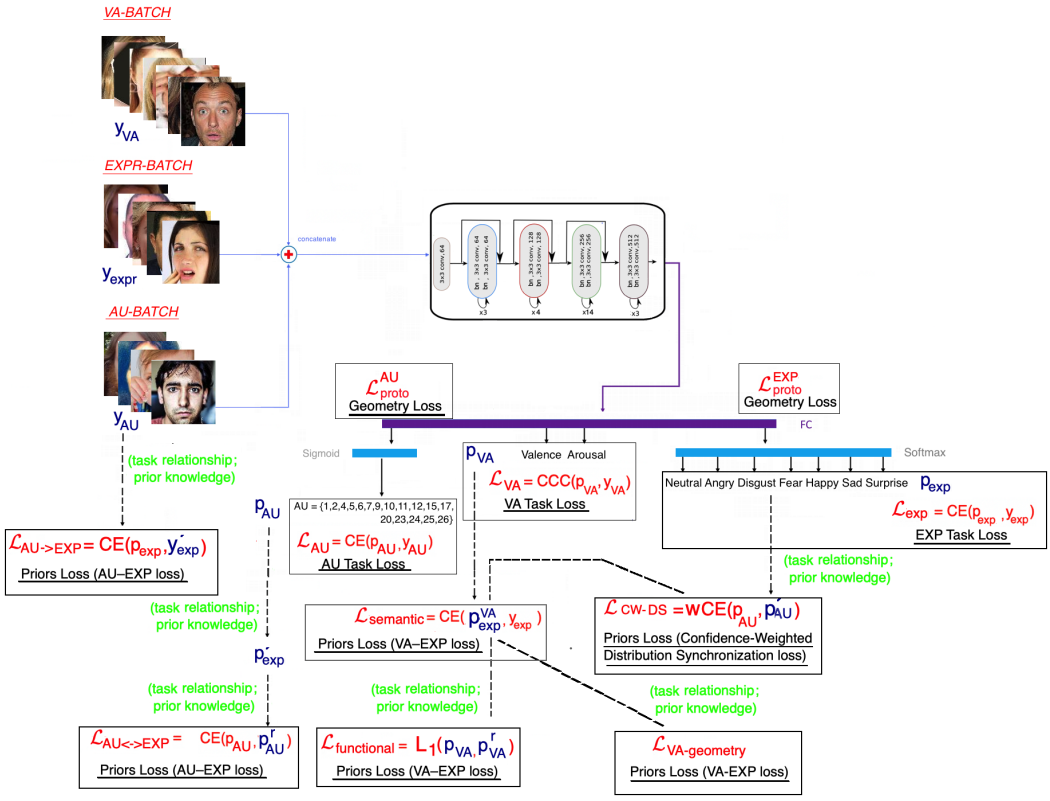

Figure 2: FaceMultiBehaviourNet integrates 3 complementary components: (i) task-specific supervision for VA-E, EXP recognition and AU-D; (ii) cross-task dependency modelling via relational priors (VA–EXP, AU–EXP, CW-DS); (iii) geometry modelling for coherent representations across tasks.

It also enables smooth transitions between EXP: e.g., a low A, moderately negative V prediction may naturally distribute probability mass between *sadness* and *fear* rather than forcing a hard assignment.

Given VA predictions (p_{VA}), we derive a VA-implied soft EXP distribution ($p^{\text{VA}}_{\text{exp}}$), by measuring proximity to predefined EXP centroids in VA space (we apply softmax over the negative squared distances between predicted VA and EXP centroids of eq. 1):

$$p^{\text{VA}}_{\text{exp}} = \text{softmax}_{exp}\left(-\|p_{\text{VA}} - c_{exp}\|^2/T\right). \tag{3}$$

The model is encouraged to assign high probability to EXP whose centroids are close to the predicted VA and low probability otherwise:

$$\mathcal{L}_{\text{semantic}} = -\sum_{j=1}^{M}\sum_{exp} y^{j}_{exp} \cdot \log(p^{\text{VA},\,j}_{exp}) \tag{4}$$

$\mathcal{L}_{\textit{VA-geometry}}$*:* The previous term does not explicitly structure the geometry of the VA manifold. Thus this term imposes angular separation between EXP directions in VA manifold, encouraging distinct, well-structured affective regions. In affective representation, the angle of a VA vector corresponds to emotional type (e.g., happiness vs. anger), whereas its magnitude reflects intensity. To respect this geometry, we adopt this term, where the angle between a normalized VA embedding and its corresponding EXP centroid is scaled by a factor α and shifted by a fixed margin m. This margin widens the angular distance between EXP directions, producing sharper decision boundaries and enhanced discriminability within the VA space. EXP such as *happiness* and *anger*, which may share high A but differ in V, are forced to occupy well-separated directions.

The predicted normalized VA vector is: $\hat{p}_{\text{VA}} = \frac{p_{\text{VA}}}{\max(\|p_{\text{VA}}\|_2, \epsilon)}$; the normalized VA centroid of EXP is: $\hat{c}_{exp} = \frac{c_{exp}}{\max(\|c_{exp}\|_2, \epsilon)}$. Using $p^{\text{VA}}_{\text{exp}}$ of eq. 3, we define a weighted target direction: $\hat{u} = \frac{u}{\max(\|u\|_2, \epsilon)}$,

where: $u = \sum_{exp} p_{\text{exp}}^{\text{VA}}\, \hat{c}_{exp}$. We impose an angular-margin constraint between $\hat{p}_{\text{VA}}$ & $\hat{u}$:

$$\mathcal{L}_{\text{VA-geometry}} = -\log \frac{e^{\alpha \cdot \cos(\theta_{\hat{u}} + m)}}{e^{\alpha \cdot \cos(\theta_{\hat{u}} + m)} + \sum_{exp} e^{\alpha \cdot \cos \theta_{\hat{c}_{exp}}}}, \tag{5}$$

α is scaling factor controlling softmax sharpness; m is additive angular margin enlarging inter-EXP separation; $\theta_{\hat{u}} = \arccos(\hat{p}_{\text{VA}}^{\top} \cdot \hat{u})$ is $\hat{p}_{\text{VA}}$ and $\hat{u}$ angle; $\theta_{\hat{c}_{exp}} = \arccos(\hat{p}_{\text{VA}}^{\top} \cdot \hat{c}_{exp})$ is $\hat{p}_{\text{VA}}$, $\hat{c}_{exp}$ angle.

$\mathcal{L}_{functional}$: While the semantic term teaches how VA implies EXP, it does not ensure the reverse, i.e., whether the inferred EXP corresponds to the same region in VA space. This asymmetry can cause semantic drift, where the two heads perform well individually but disagree conceptually (e.g., predicting "surprise" should not reconstruct to a low A point). This term closes this loop by enforcing functional consistency between the two representations, ensuring that VA and EXP form a coherent, invertible mapping rather than a one-way association.

We map $p_{\text{exp}}^{\text{VA}}$ of eq. 3 back to VA space using the centroids of eq. 1, where p_{VA}^{r} is reconstructed VA estimate: $p_{\text{VA}}^{r} = \sum_{exp} p_{\text{exp}}^{\text{VA}}\, c_{exp}$, required to match the original prediction:

$$\mathcal{L}_{\text{functional}} = \|p_{\text{VA}} - p_{\text{VA}}^{r}\|_1. \tag{6}$$

AU–EXP loss EXPs emerge from structured patterns of facial muscle activations. AUs thus provide fine-grained and physically grounded description of facial behaviour, whereas EXPs provide higher-level semantic abstraction. This makes AU information valuable not only as auxiliary signal, but also as source of structured supervision for EXP learning when EXP labels are missing, or noisy. Naïve MTL may cause AU and EXP heads to drift toward semantically inconsistent predictions, e.g., predicting "happy" while suppressing AU12. To prevent this, AU and EXP predictions should be mutually consistent: if an AU set implies a particular EXP, then predicting that EXP should in turn imply a compatible AU configuration. Enforcing this bidirectional relation reduces semantic drift, stabilises learning under partial supervision, improves generalisation across heterogeneous labels.

This loss jointly enforces (i) *bottom-up supervision* from AUs to EXP, (ii) *top-down and bottom-up self-consistency* between AU, EXP. It combines 2 complementary terms operating at different levels but sharing common goal: aligning low-level muscular evidence with high-level categorical affect.

$$\mathcal{L}_{\text{AU–EXP}} = \lambda_{\text{r}} \mathcal{L}_{\text{AU}\to\text{EXP}} + \lambda_{\text{c}} \mathcal{L}_{\text{AU}\leftrightarrow\text{EXP}}. \tag{7}$$

$\mathcal{L}_{AU\to EXP}$: It transfers supervision from AUs to EXP by constructing soft EXP targets from AU evidence, allowing ambiguous AU configurations to map to graded EXP probs rather than forced one-hot labels, e.g., AU6 and AU12 strongly support *happy*, whereas AU4 and AU15 yield softer mass over *sad*-related EXPs. Thus, EXP learning remains stable under missing or noisy EXP annotations.

We compute an EXP-related score S_{exp} and from that we derive soft EXP labels y'_{exp}, as follows:

$$S_{exp} = \frac{\sum_{\text{AU}_i} w_{exp,i}\, y_{\text{AU}_i}}{\sum_{\text{AU}_i} w_{exp,i}}, \qquad y'_{exp} = \text{softmax}(S_{exp}), \tag{8}$$

where: w_{AU_i} is weight reflecting the association strength of *i*-th AU & EXP; it is 1 if AU_i is prototypical for EXP, w if observational, 0 otherwise. The resulting soft labels supervise EXP head:

$$\mathcal{L}_{\text{AU}\to\text{EXP}} = -\sum_{j=1}^{M} \sum_{exp} y'^{\,j}_{exp} \log(p^{j}_{exp}). \tag{9}$$

$\mathcal{L}_{AU\leftrightarrow EXP}$: It enforces consistency between AU and EXP by requiring that AU-implied EXP, when mapped back, reproduce the original AU configuration. It prevents semantic drift between task heads by penalising incompatible predictions, e.g., predicting "happy" without activating its characteristic AUs. The constraint applies even in absence of AU labels, acting as regulariser during joint training.

From predicted AUs p_{AU}, we infer a soft EXP prediction p'_{exp} as follows:

$$s'_{\text{exp}} = \frac{\sum_{\text{AU}_i} w_{\text{AU}_i} \cdot p_{\text{AU}_i}}{\sum_{\text{AU}_i} w_{\text{AU}_i}}, \quad p'_{\text{exp}} = \text{softmax}(s'_{\text{exp}}), \tag{10}$$

where: w_{AU_i} is 1 if AU_i is prototypical for the EXP, w if it is observational and 0 otherwise.

We then map the soft EXP predictions back to reconstructed AUs: $p^{\text{r}}_{\text{AU}_i} = \sum_{\text{exp}} p'_{\text{exp}} \cdot p_{\text{AU}_i|\text{exp}}$. Finally, we compare the reconstructed AUs with the original AU predictions p_{AU}:

$$\mathcal{L}_{\text{AU}\leftrightarrow\text{EXP}} = -\sum_{j=1}^{M} \sum_{\text{AU}_i} p^{r,j}_{\text{AU}_i} \cdot \log(p^{j}_{\text{AU}_i}). \tag{11}$$

Confidence-Weighted Distribution Synchronization loss The previous two loss terms do not explicitly regulate how strongly cross-task constraints should be applied at the *instance level*. In practice, affect predictions may be ambiguous for certain samples (e.g., neutral faces or transitional expressions) and enforcing strong cross-task alignment in such cases can introduce noisy gradients and destabilise training. EXP and AU heads may become statistically misaligned during joint training, especially under class imbalance or partial supervision, e.g., EXP head may predict *happy* while AU head suppresses AU12, simply because each head optimises its own objective. However, enforcing strict alignment for all samples is undesirable: when affective state is inherently ambiguous (e.g., low A, near-neutral V), forcing agreement can amplify noise rather than improve coherence. Thus, this term synchronises AU and EXP predictions while adaptively modulating its strength based on model's confidence in the underlying affective state. Unlike the previous losses, which impose structural constraints regardless of certainty, this term acts as a *trust-aware regulariser*, enforcing cross-task agreement only when the affect estimate is reliable, reducing the impact of noisy or ambiguous samples, mitigating conflicting gradients, stabilising optimisation in heterogeneous settings.

From predicted EXP p_{exp}, we derive soft AU targets via $p_{\text{AU}_i|\text{exp}}$: $p'_{\text{AU}_i} = \sum_{\text{exp}} p_{\text{exp}} \cdot p_{\text{AU}_i|\text{exp}}$. Then, for each sample, we compute its entropy based on $p^{\text{VA}}_{\text{exp}}$ of eq. 3: $H_{\text{VA}} = -\sum_{exp} p^{\text{VA}}_{\text{exp}} \log p^{\text{VA}}_{\text{exp}}$. Then we derive a confidence weight: $w_{\text{VA}} = 1 - H_{\text{VA}} / \log 7 \in [0, 1]$. When VA confidence is high (e.g., strong positive V, moderate A), w_{VA} is large, enforcing tight alignment between EXP-driven AU expectations and AU predictions, e.g., *happy* predictions are encouraged to activate AU6 and AU12. When VA confidence is low (e.g., near-neutral, mixed affect), the loss is attenuated, preventing uncertain VA–EXP mappings from imposing incorrect AU constraints. This yields a trust-adaptive coupling mechanism that promotes cross-task consistency only when affective state is well defined:

$$\mathcal{L}_{\text{CW-DS}} = -\lambda_{cw-ds} \sum_{j=1}^{M} \sum_{\text{AU}_i} w^{j}_{\text{VA}} \cdot p'^{j}_{\text{AU}_i} \cdot \log(p^{j}_{\text{AU}_i}). \tag{12}$$

Geometry Loss In a well-structured affective space, related concepts across tasks should occupy nearby regions of the embedding manifold, eg., samples predicted as *happy* should lie close to EXP prototype "happy" and AU6 and AU12 prototypes. Conversely, samples associated with *sad* should be geometrically distant from "happy" and its AUs. In MTL, task-specific heads can implicitly induce incompatible geometric organisations, even when their predictions are mutually consistent at label or distribution level. This misalignment weakens feature sharing, limiting cross-task transfer. This loss addresses this by explicitly aligning the geometry induced by different task heads. It operates directly on classifier prototypes (i.e., head weight vectors) and encourages samples to align with semantically compatible directions across tasks, while remaining separated from incompatible ones.

Relation to $\mathcal{L}_{\text{VA-geometry}}$ While $\mathcal{L}_{\text{VA-geometry}}$ structures geometry *within VA manifold* by separating EXP directions, this loss operates in *shared latent feature space* and jointly considers all tasks. The former enforces inter-EXP separation in continuous affect space; the latter enforces cross-task coherence & intra-class compactness across discrete and continuous tasks. Together, they provide complementary geometric constraints: one shapes affective directions, the other aligns task-specific decision boundaries.

Each task head terminates in a linear classifier with weight vectors that act as class prototypes. Let $\mathbf{W}_{\text{EXP}} = \{\mathbf{w}_{exp}\}$ denote EXP prototypes, $\mathbf{W}_{\text{AU}_i} = \{\mathbf{w}^{+}_{\text{AU}_i}, \mathbf{w}^{-}_{\text{AU}_i}\}$ denote positive and negative AU prototypes. Given the shared embedding $\mathbf{z}$ (producing task-specific predictions), we define contrastive objectives that maximise similarity to correct prototypes and minimise similarity to incorrect ones:

$$\mathcal{L}^{\text{EXP}}_{\text{proto}} = -\log \frac{e^{\text{sim}(\mathbf{z}, \mathbf{w}_{exp})/\tau}}{\sum_{exp'} e^{\text{sim}(\mathbf{z}, \mathbf{w}_{exp'})/\tau}}, \tag{13}$$

$$\mathcal{L}^{\text{AU}}_{\text{proto}} = -\log \frac{e^{\text{sim}(\mathbf{z}, \mathbf{w}^{s}_{\text{AU}_i})/\tau}}{\sum_{s' \in \{+,-\}} e^{\text{sim}(\mathbf{z}, \mathbf{w}^{s'}_{\text{AU}_i})/\tau}}, \tag{14}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity, $s \in \{+, -\}$ depends on AU activation, τ is temperature parameter.

4 Experimental Results

Training Implementation Details For training MultiFaceBehaviorNet, we use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay 1×10^{-4}. The base learning rate is 5×10^{-4}, with a 5-epoch linear warm-up followed by cosine decay down to 1×10^{-6}. Training runs for 50 epochs with a batch size of 128. We apply gradient clipping with $\|\nabla\|_2 \leq 1.0$. The input images to MultiFaceBehaviorNet are resized to dimensions $112 \times 112 \times 3$ and the pixel intensity values to $[-1, 1]$. Training was performed on a Tesla V100 32GB GPU; training time was about 2 days. Unless otherwise stated, all non-task losses use fixed coefficients after an initial ramp-up. The final weights are: $\lambda_{\text{VA}} = 1.0$; $\lambda_{\text{EXP}} = 1.0$; $\lambda_{\text{AU}} = 1.0$; $\lambda_s = 0.1$; $\lambda_g = 0.05$; $\lambda_f = 0.1$; $\lambda_r = 0.05$; $\lambda_c = 0.05$; $\lambda_{cw-ds} = 0.1$. These values balance cross-task coupling while preventing any single regulariser from dominating training.

To avoid gradient shocks, conflicting updates and negative transfer (especially when combining heterogeneous datasets) the full objective is introduced progressively across the following stages: (i) *Phase 1 (Task-only warm-up; epochs 1-6):* Only task losses are active, allowing backbone and heads to stabilise before introducing dependencies; (ii) *Phase 2 (Prior-based coupling; epochs 7-27):* We ramp them to full strength. This injects structured task relationships while maintaining stability; *Phase 3 (Geometry shaping; epochs 28-50):* We ramp them during this phase's early epochs. This shapes a coherent, discriminative affect manifold shared by all tasks and boosts cross-db generalisation.

Implementation details, experimental results for face detection and localisation are in appendix.

Datasets Aff-Wild [25] is annotated for VA, is split into training and test sets of 252 (1M frames) and 46 videos (220K frames). **AffectNet** [43] is annotated for BER and VA, is split into training and test sets of 280K and 3.5K images. **RAF-DB** [37] is annotated for BER, is split into training and test sets of 12K and 3K images. **EmotioNet** [11] is annotated for 11 AUs, is split into training and validation sets, each with 25K images. **DISFA** [42] is annotated for 12 AUs, consists of 27 videos, each with 5K frames. **BP4D-Spontaneous** [66] (BP4D) is annotated for 12 AUs, consists of 41 subjects of 140K frames. **BP4D+** [68] is annotated for 12 AUs, consists of 140 subjects of 190K frames. In our experiments, we merge these datasets' training sets and evaluate on each dataset's test set. More details on each db exist in appendix.

Recent studies [13, 14] have highlighted inconsistencies in dataset partitioning and evaluation practices in ABA, leading to unfair comparisons. To address this, a unified partitioning protocol was proposed, incorporating demographic information to ensure fairness and comparability. It was shown that methods previously considered sota may not perform as well under this new protocol. In this work, we further adopt this partitioning protocol for AffectNet, RAF-DB, EmotioNet, and DISFA, which were re-annotated and repartitioned (referred to as 'New' in our experiments). Aff-Wild, BP4D and BP4D+ do not contain demographic information and have not been restructured by [13, 14].

Performance Measures For overall performance, we use: i) average CCC of VA for evaluating VA-E on Aff-Wild, AffectNet; CCC takes values in $[-1, 1]$; ii) average F1 across all EXP for evaluating EXP recognition on RAF-DB, AffectNet; iii) average F1 across all AUs for evaluating AU-D on DISFA, EmotioNet, BP4D, BP4D+; F1 score takes values in $[0, 1]$. Higher values are desirable for all metrics. We use F1, instead of accuracy, e.g., which is standard in RAF-DB, as it is a stricter and more informative metric for imbalanced datasets. In fact, high accuracy can be achieved by simply predicting the most frequent classes. More details about all evaluation metrics exist in appendix.

4.1 Overall & Fairness Evaluation

BER Table 3 provides performance comparison (in terms of F1) for BER on AffectNet and RAF-DB. The comparison includes our proposed toolkit, the sota methods and other existing toolkits. When utilizing the original dataset splits for model training and evaluation, our toolkit demonstrates a superior performance exceeding the sota by 10.7% on RAF-DB. Moreover, our toolkit significantly outperforms the other toolkits, with improvements of at least 13.1% on AffectNet and 20.2% on RAF-DB. When adopting the new dataset splits proposed by [14, 13], our toolkit surpasses the sota by at least 2.6% on AffectNet and 4.8% on RAF-DB. Notably, the performance gains over the sota are even more pronounced when using the new dataset splits compared to the original splits.

VA-E Table 4 provides a performance comparison (in terms of CCC) for VA Estimation on Aff-Wild and AffectNet. The comparison includes our proposed toolkit and the sota methods (since no existing toolkit performs this task). When utilizing the original dataset splits for model training and evaluation,

our toolkit demonstrates a superior performance, exceeding sota by at least 6.2% both on Aff-Wild and AffectNet. When utilizing the new split for AffectNet [14, 13], still exceeds sota by at least 5.3%.

AUD Table 5 provides performance comparison (in terms of F1) for AU Detection on EmotioNet, DISFA, BP4D and BP4D+. The comparison includes our proposed toolkit, the sota methods and other existing toolkits. When utilizing the original dataset splits for model training and evaluation, our toolkit demonstrates a superior performance, exceeding the sota by at least 18.2% on EmotioNet, 12% on DISFA, 17.9% on BP4D and 8.4% on BP4D+. Moreover, our toolkit significantly outperforms the other toolkits, with improvements of over 45.3% on EmotioNet, 19.1% on DISFA, 23% on BP4D and 19.6% on BP4D+. When adopting the new dataset splits proposed by [14, 13], our toolkit continues to exhibit leading performance, surpassing the sota by at least 8.7% on EmotioNet and 22% on DISFA.

Table 3: BER: F1 (in %) (↑)

Datasets	AffectNet		RAF-DB	
	Original	New	Original	New
DAN	65.7	60.0	64.5	70.9
EAC	65.3	60.3	63.6	75.5
MA-Net	64.5	55.4	60.7	65.2
EffFace	63.7	58.6	68.7	73.2
POSTER++	67.4	59.6	68.9	73.7
UA-FER	66.9	58.9	68.8	72.3
MHAN	67.7	59.2	67.7	72.5
CLIPER	66.2	58.1	67.9	72.1
CLIP	64.7	55.9	66.6	71.5
Norface	**68.6**	60.2	69.1	74.5
py-feat	55.0	-	49.3	-
LibreFace	49.7	-	59.6	-
Ours	68.1	**62.9**	**79.8**	**80.3**

Table 4: VA-E: CCC (in %) (↑)

Datasets	Aff-Wild	AffectNet	
	Original	Original	New
FUXI	52.0	56.5	74.0
SITU	53.1	57.5	71.1
CTC	50.2	52.3	71.0
USTC-IAT	51.4	53.7	71.2
GRJCA	50.9	53.9	71.6
CtyunAI	50.6	53.5	71.2
VA-StarGan	50.2	54.5	72.4
MT-EffNet	51.7	57.2	71.8
Ours	**59.3**	**63.7**	**79.3**

Table 5: AUD: F1 (in %) (↑)

Datasets	EmotioNet		DISFA		BP4D	BP4D+
	Orig	New	Orig	New	Orig	Orig
FUXI	60.2	77.9	49.8	43.7	63.2	56.2
SITU	59.7	77.2	49.2	40.4	62.7	55.8
CLEF	57.6	74.6	64.8	45.9	65.9	63.1
AUFormer	63.6	78.8	66.4	46.2	66.2	62.7
FMAE-IAT	59.8	72.9	70.1	46.1	67.1	66.8
MDHR	60.6	77.2	64.9	45.1	66.6	63.4
MCM	58.6	76.4	64.3	43.8	66.0	62.7
py-feat	36.5	-	54.0	-	61.4	52.4
LibreFace	35.7	-	63.0	-	62.0	55.6
OpenFace	35.5	-	59.0	-	61.3	53.5
Ours	**81.8**	**87.5**	**82.1**	**68.2**	**85.0**	**75.2**

Fairness We further evaluate Behaviour4All with respect to demographic fairness, including age, gender, race. The detailed results are in appendix. Overall, Behaviour4All consistently achieves better fairness than both sota and existing toolkits across all tasks and datasets. These results indicate that the proposed dependency-aware MTL framework not only improves overall predictive performance, but also leads to more balanced behaviour across demographic groups.

Finally, one could assume that our **toolkit's sota performance primarily results from the increased training data amount, rather than our method's various components**. However, this is not true as we show in appendix, reinforcing prior findings [59, 63] that incorporating more (same-task) data is detrimental, as labeling inconsistencies, annotation errors & biases are present.

4.2 Ablations, Generalizability, Downstream Tasks & Efficiency

Ablation Studies Table 6 compares: single-task (ST); vanilla MTL (only task losses); MTL with task losses + geometry; MTL with task losses + relational priors; the full framework. For vanilla MTL, negative transfer appears for BER, where ST outperforms MT on all EXP datasets. Introducing relational priors improves performance across all datasets. Adding geometry modelling further improves performance. Overall, the full model consistently achieves the best results and alleviates negative transfer. In appendix, we further conduct an ablation indicating the impact of each proposed loss term and compare against a Student–Teacher MTL baseline based on knowledge distillation and pseudo-label completion. The latter still suffers from negative transfer (for EXPs), while our framework consistently achieves substantially better performance across all datasets and tasks.

Table 6: Ablation studies analysing: (i) single-task (ST) vs. multitask (MT) learning; and (ii) the contribution of each principled component of the proposed dependency-aware MTL framework.

Setting	Task	Priors	Geometry	Aff-Wild	Aff-VA	Aff-BE	RAF-DB	EmotioNet	DISFA	BP4D	BP4D+
ST	–	–	–	53.3	58.1	64.3	68.4	51.9	59.9	79.9	58.1
MT	✓	✗	✗	55.8	59.2	62.1	65.1	53.1	61.1	81.1	59.6
MT + Geometry	✓	✗	✓	56.5	59.7	62.7	67.4	56.9	62.8	81.1	59.7
MT + Priors	✓	✓	✗	58.3	63.0	67.1	77.1	76.8	80.1	85.0	75.0
Full	✓	✓	✓	**59.3**	**63.7**	**68.1**	**79.8**	**81.8**	**82.1**	**85.0**	**75.2**

Generalisability The exceptional generalization of our toolkit across the test sets of the 8 datasets used in its training serves as a strong indicator of its effectiveness and versatility. To further demonstrate and validate the robustness and quality of its learned features, we show that it can generalize its knowledge & capabilities to unseen datasets not utilized in its training that possess different statistical properties and contexts. Table 7(a) provides performance comparison for BER on Aff-Wild2 [26, 27, 23] (w.r.t. F1), for AUD on GFT [12] (w.r.t. F1), for VA-E on AFEW-VA [34] (w.r.t. CCC). The comparison includes our toolkit, the sota methods (AffWildNet, JAA-NET, FUXI) -that have been trained on one of these datasets-, and an existing toolkit (AFFDEX 2.0). Our toolkit outperforms FUXI by 7%, AFFDEX 2.0 by 5.5%, AffWildNet by 16.5%, JAA-Net by 8%.

Downstream Task: Compound Expression Recognition (CER) We extend our experiments by using the proposed toolkit as foundation model, leveraging its rich facial behaviour representations for efficient transfer learning. We evaluate its zero-shot and few-shot performance on CER to assess adaptability with minimal or no task-specific data. These experiments demonstrate strong generalization, robustness to new tasks, and reduced reliance on large labeled datasets. More details exist in appendix. In Table 7(b) it can be seen that, on EmotioNet and RAF-DB, our toolkit consistently outperforms all sota by significant margins in both zero- and few-shot settings.

Table 7: Generalisation and downstream task evaluation (in %). AA denotes average accuracy.

(a) Cross-dataset generalisation

Datasets	Aff-Wild2-Expr F1	AFEW-VA CCC	GFT F1
JAA-Net	–	–	55.0
AffWildNet	–	54.0	–
FUXI	34.5	–	–
AFFDEX 2.0	36.0	–	–
Behaviour4All	**41.5**	**70.5**	**63.0**

(b) Compound EXP Recognition

Datasets	EmotioNet F1	RAF-DB AA
NTechLab	25.5	–
VGG-FACE-mSVM	–	31.6
DLP-CNN	–	44.6
zero-shot Behaviour4All	**32.9**	**47.9**
fine-tuned Behaviour4All	**40.7**	**57.5**

Efficiency In appendix, we compare the computation cost of our toolkit (for predicting the 3 behaviour tasks) to other toolkits w.r.t. FPS, total number of parameters, GFLOPs. Our toolkit is of almost similar efficiency to LibreFace. Our toolkit provides more accurate VA, EXP, AU analysis whilst running much faster than OpenFace, py-feat. We highlight that our reported efficiency metrics reflect the performance of our toolkit when simultaneously predicting VA, AUs, EXP, whereas LibreFace and py-feat predict only AUs and EXP concurrently; OpenFace is limited to predicting AUs only.

Optimisation dynamics and gradient conflict analysis To further analyse the optimisation behaviour of Behaviour4All, the supplementary material include: (i) an analysis of the staged optimisation dynamics of all losses during training, and (ii) an explicit gradient-level study comparing vanilla MTL against our framework. The analysis shows that the proposed formulation progressively improves gradient alignment between tasks, substantially reduces gradient conflict and leads to stable joint optimisation.

5 Conclusions

We presented Behaviour4All, a dependency-aware MTL framework for facial behaviour analysis. By combining task-specific supervision, relational priors and geometry modelling, our framework explicitly captures dependencies between VA, EXP & AUs, mitigating negative transfer and supporting learning under heterogeneous annotations. Experiments across 13 in-the-wild datasets show sota performance, improved fairness, strong generalisation. Additional supplementary analyses demonstrate stable optimisation, reduced gradient conflict and applicability beyond affective computing.

Acknowledgments and Disclosure of Funding

The work of Chunchang Shao has been supported by a Principal Scholarship of Queen Mary University of London.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[2] C.F. Benitez-Quiroz, R. Srinivasan, and A.M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16)*, Las Vegas, NV, USA, June 2016.

[3] Mina Bishay, Kenneth Preston, Matthew Strafuss, Graham Page, Jay Turcot, and Mohammad Mavadati. Affdex 2.0: A real-time facial expression analysis toolkit. In *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2023.

[4] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 354–370. Springer, 2016.

[5] Di Chang, Yufeng Yin, Zongjian Li, Minh Tran, and Mohammad Soleymani. Libreface: An open-source toolkit for deep facial expression analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8205–8215, 2024.

[6] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *European conference on computer vision*, pages 609–623. Springer, 2012.

[7] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J Chang. Py-feat: Python facial expression analysis toolbox. *Affective Science*, 4(4):781–796, 2023.

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[9] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[10] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[11] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[12] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 581–588. IEEE, 2017.

[13] Guanyu Hu, Dimitrios Kollias, Eleni Papadopoulou, Paraskevi Tzouveli, Jie Wei, and Xinyu Yang. Rethinking affect analysis: A protocol for ensuring fairness and consistency. *arXiv preprint arXiv:2408.02164*, 2024.

[14] Guanyu Hu, Eleni Papadopoulou, Dimitrios Kollias, Paraskevi Tzouveli, Jie Wei, and Xinyu Yang. Bridging the gap: Protocol towards fair and consistent affect analysis. *arXiv preprint arXiv:2405.06841*, 2024.

[15] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017.

[16] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800.

[17] Dimitrios Kollias. 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*, pages 2328–2336, 2022.

[18] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023.

[19] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, pages 1–30, 2020.

[20] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[21] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. 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*, pages 5888–5897, 2023.

[22] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Simon Bacon, Alice Baird, Chris Gagne, et al. Advancements in affective and behavior analysis: The 8th abaw workshop and competition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5572–5583, 2025.

[23] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Simon Bacon, Jens Madsen, Soufiane Belharbi, Muhammad Haseeb Aslam, Chunchang Shao, and Guanyu Hu. From affect to complex behavior: Advancing multimodal human-centered ai at the 10th abaw workshop & competition, 2026.

[24] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024.

[25] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907–929, 2019.

[26] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[27] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[28] Dimitrios Kollias and Stefanos Zafeiriou. Va-stargan: Continuous affect generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 227–238. Springer, 2020.

[29] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[30] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021.

[31] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Abhinav Dhall, Shreya Ghosh, Chunchang Shao, and Guanyu Hu. 7th abaw competition: Multi-task learning and compound expression recognition. In *European Conference on Computer Vision*, pages 31–45. Springer, 2024.

[32] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Eric Granger, Marco Pedersoli, Alessandro Lameiras Koerich, Simon L Bacon, Oya Celiktutan, Soufiane Belharbi, Muhammad Osama Zeeshan, et al. From complex behavior to intelligent human-centered ai: The 11th abaw workshop & competition. In *11th Workshop and Competition on Affective {\&} Behavior Analysis in-the-wild*.

[33] Dimitrios Kollias, Stefanos Zafeiriou, Irene Kotsia, Greg Slabaugh, Damith Chamalke Senadeera, Jianian Zheng, Kaushal Kumar Keshlal Yadav, Chunchang Shao, and Guanyu Hu. From emotions to violence: Multimodal fine-grained behavior analysis at the 9th abaw. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–12, 2025.

[34] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[35] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980. IEEE, 2012.

[36] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[37] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 2584–2593. IEEE, 2017.

[38] Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, and Wenqiang Jiang. Facial expression recognition based on multi-modal features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5878, 2023.

[39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[40] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1239–1246, 2022.

[41] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, and Yigang Wang. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 157:110951, 2025.

[42] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013.

[43] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[44] Hajime Nada, Vishwanath A Sindagi, He Zhang, and Vishal M Patel. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–10. IEEE, 2018.

[45] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry Davis. SSH: Single stage headless face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[46] Mang Ning, Albert Ali Salah, and Itir Onal Ertugrul. Representation learning and identity adversarial training for facial behavior understanding. *arXiv preprint arXiv:2407.11243*, 2024.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[48] Gnana Praveen Rajasekhar, Jahangir Alam, and Eric Charton. United we stand, divided we fall: Handling weak complementarity for audio-visual emotion recognition in valence-arousal space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5741–5751, 2025.

[49] Andrés Romero, Juan León, and Pablo Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 2018.

[50] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[51] Andrey V Savchenko. Mt-emotieffnet for multi-task human affective behavior analysis and learning from synthetic data. In *European Conference on Computer Vision*, pages 45–59. Springer, 2022.

[52] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*, 2017.

[53] Xiaofeng Wang, Tianbo Han, Songling Liu, Muhammad Shahroz Ajmal, Lu Chen, Yongqin Zhang, and Yonghuai Liu. Mhan: Multi-head hybrid attention network for facial expression recognition. *Pattern Recognition*, page 112015, 2025.

[54] Zihan Wang, Siyang Song, Cheng Luo, Songhe Deng, Weicheng Xie, and Linlin Shen. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1270–1280, 2024.

[55] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11293–11302, 2019.

[56] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2):199, 2023.

[57] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[58] Shuo Yang, Yuanjun Xiong, Chen Change Loy, and Xiaoou Tang. Face detection through scale-friendly deep convolutional networks. *arXiv preprint arXiv:1706.02863*, 2017.

[59] Chengyan Yu, Dong Zhang, Wei Zou, and Ming Li. Joint training on multiple datasets with inconsistent labeling criteria for facial expression recognition. *IEEE Transactions on Affective Computing*, 2024.

[60] Jun Yu, Yang Zheng, Lei Wang, Yongqi Wang, and Shengfan Xu. Cross-modal facial expression recognition with global channel-spatial attention: Modal enhancement and proportional criterion fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5707–5714, 2025.

[61] Kaishen Yuan, Zitong Yu, Xin Liu, Weicheng Xie, Huanjing Yue, and Jingyu Yang. Auformer: Vision transformers are parameter-efficient facial action unit detectors. In *European Conference on Computer Vision*, pages 427–445. Springer, 2024.

[62] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017.

[63] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.

[64] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[65] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multi-modal facial affective analysis based on masked autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2023.

[66] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[67] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision*, pages 418–434. Springer, 2022.

[68] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.

[69] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.

[70] Haoliang Zhou, Shucheng Huang, and Yuqiao Xu. Ua-fer: Uncertainty-aware representation learning for facial expression recognition. *Neurocomputing*, 621:129261, 2025.

[71] Weiwei Zhou, Chenkun Ling, and Zefeng Cai. Emotion recognition with clip and sequential learning. *arXiv preprint arXiv:2503.09929*, 2025.

[72] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. 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*, pages 5756–5763, 2023.

[73] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. *Deep learning for biometrics*, pages 57–79, 2017.

A Related Work: Toolkits and State-of-the-art

A.1 Toolkits Clarification

At first let us mention that on Table 1 of the main paper, LibreFace is denoted as not having face bounding boxes and landmarks. The authors of LibreFace state that they leverage MediaPipe, a third-party detector. LibreFace does not develop its own face detection or landmark model. Since face localization is not part of their method's contributions, we chose not to credit it as a native component in Table 1. We also note that the use of MediaPipe introduces external dependency risks - if it becomes deprecated or unavailable, LibreFace's functionality may be compromised.

A.2 State-of-the-art Works

DAN [56] consists of 3 components that enhance class separability, while focusing on multiple facial regions simultaneously. MA-Net [69] combines a multi-scale module to enhance feature diversity and robustness with a local attention module that focuses on salient facial regions. EAC [67] improves BER under noisy labels by using attention consistency combined with random erasing to prevent the model from memorizing noisy samples. POSTER++ [41] improves BER by replacing the original complex two-stream pyramid cross-fusion architecture of POSTER with a single-stream model that fuses landmark and image features via window-based cross-attention; it simplifies the multi-scale feature extraction by integrating the landmark and image features at multiple scales. UA-FER [70] integrates a vision–language pre-training backbone with evidential deep-learning to tackle BER by using a Multi-granularity Feature Decoupling module and by introducing a Relation Uncertainty Calibration module. MHAN [53] is a lightweight CNN for BER consisting of an Efficient Local Attention Mixed Feature Network and a Multi-Head Hybrid Attention mechanism (that overlays multiple attention heads to generate diverse attention maps). CLIPER [36] leverages a frozen CLIP by learning multiple fine-grained learnable text descriptors for each expression, then training in two stages (first learning the text descriptors while keeping encoders frozen, then fine-tuning the image encoder with those descriptors fixed).

ME-GraphAU [40] employs a multi-dimensional edge feature-based AU relation graph that learns the relationships between pairs of AUs. AUNets [49] predicts the viewpoint of a video first and then applies an ensemble of AU detectors specifically trained for that viewpoint. Res50, the winner of Emotionet Challenge, employed a ResNet50 and extra automatically annotated images. FMAE-IAT[46] first pre-trains a facial Vision Transformer backbone via a masked-autoencoder architecture, and then fine-tunes it for AU-D using a dual-head scheme where an AU prediction head is trained to minimize AU loss while concurrently an identity prediction head is trained adversarially. AUFormer [61] freezes a pre-trained ViT backbone, then inserts a set of lightweight "Mixture-of-Knowledge Expert" modules and then allows them to collaborate within their expert group by averaging their knowledge and injecting that aggregated information back into the frozen ViT blocks, while simultaneously training with a margin-truncated difficulty-aware weighted asymmetric loss. MDHR [54] is a model that captures facial motion at multiple spatial scales with adaptive weighting and models hierarchical spatial-temporal relationships among AUs, first locally (within facial regions) then cross-regionally.

AffWildNet [26] is an end-to-end CNN-RNN, which integrates facial landmarks within the network's design, trained with a correlation-based loss and employing a rebalancing data augmentation strategy. VA-StarGan is a VGG16 trained with both real and generated images of various VA states. MT-EffNet [51] is a MTL framework leveraging EfficientNet as a backbone to jointly learn BE, AUs, and VA.

FUXI uses a pre-trained MAE to extract and fine-tune visual features and then combines these with audio features using a Temporal and Multi-modal Fusion component within a Transformer-based method. SITU extracts multi-modal features with pre-trained models and combines them with temporal encoders to capture contextual information. CTC uses a combination of TCN and Transformers to fuse visual & audio features. GRJCA[48] proposes a Gated Recursive Joint Cross-Attention mechanism that adaptively controls how much to rely on cross-modal attended features vs. original features across multiple iterations. CtyunAI[71] first fine-tune a visual encoder (based on CLIP) and then feed the extracted features into a TCN and a Transformer encoder to capture temporal dynamics and perform affect recognition. FUXI [65], SITU [38], CTC [72] are the top and best performing methods on the ABAW Competition for BER, AUD & VA-E.

B Behaviour4All Toolkit

B.1 FaceLocalizationNet

For face detection and alignment we used RetinaFace [8], which is a state-of-the-art deep learning-based face detector designed for robust and high-accuracy face detection. The key idea behind RetinaFace is to detect faces with a high level of precision, even under challenging conditions such as variations in lighting, pose, occlusion, and scale. RetinaFace is a single-stage face detector, meaning that it predicts face locations and key points directly from the image without requiring a two-step process (e.g., region proposal and refinement). This makes it faster and more efficient compared to two-stage detectors like Faster R-CNN. RetinaFace consists of three main components: the feature pyramid network, the context head module and the cascade multi-task loss.

First, the feature pyramid network gets the input face images and outputs five feature maps of different scale so as to enable the detection of both large and small faces. It uses a backbone network to extract features from the input images, which is typically a ResNet. In our case, we employ both a MobileNet model, as well as a ResNet model. Then, the context head module gets a feature map of a particular scale and calculates a multi-task loss. The context head module helps in capturing more contextual information around faces, improving the detection performance in cluttered environments. RetinaFace simultaneously detects face bounding boxes, 5 (key) facial landmarks (two eyes, nose tip, and mouth corners), and provides a 3D position estimation. The multi-task approach helps in improving the accuracy of face detection by leveraging related tasks. The landmark localization branch (that predicts five facial landmarks) is particularly useful for face alignment and behaviour analysis tasks, as we will explain in a bit.

B.2 FaceMultiBehaviourNet

The FaceMultiBehaviourNet consists of a backbone network, a deep CNN composed of several residual blocks with increasing feature map sizes. It begins with a single 3×3 convolutional layer with 64 filters, followed by four stages of residual blocks. Each block contains two 3×3 convolutional layers followed by batch normalization, repeated multiple times: 3 times for 64 filters, 4 times for 128 filters, 14 times for 256 filters, and 3 times for 512 filters. The arrows indicate residual connections within and across the stages, resembling a ResNet-like structure designed for robust feature extraction. Figure 3 shows this backbone architecture.

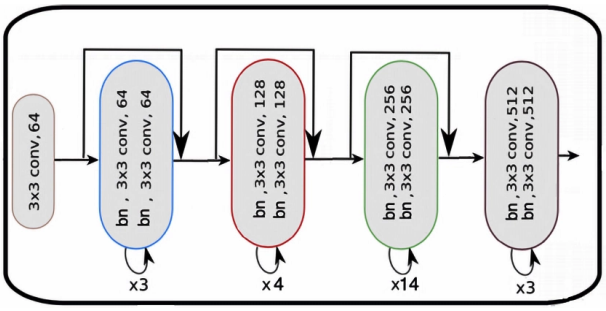

Figure 3: FaceMultiBehaviourNet's backbone architecture

C Datasets Utilized

Here, we provide more details on the datasets that have been utilized for the development (training and testing) of Behaviour4All toolkit.

The **Aff-Wild** dataset [25, 62] has been the first large scale captured in-the-wild dataset, containing 298 videos (200 subjects) of 1.25M frames, annotated in terms of valence-arousal that range in $[-1, 1]$. It served as benchmark for the Aff-Wild Challenge organized in CVPR 2017. The training set consists of 1M frames and the test set of 250K.

The **AFEW-VA** dataset [34] consists of 600 video clips (without audio) totalling to around 30K frames that were annotated in terms of valence-arousal; the values are discrete in the range of [-10,10]. The values have been scaled to [-1,1] so as to be consistent with the other datasets.

The **AffectNet** dataset [43] contains around 400K images manually annotated for 7 basic expressions (plus contempt) and VA that ranges in $[-1, 1]$. The training set of this dataset consists of around 321K images and the validation of 5K. The validation set is balanced across the different expression categories. In this work, we exclude the contempt class.

The **RAF-DB** dataset [37] contains 12.2K training and 3K test facial images annotated in terms of the 7 basic expressions. It further contains a small subset of around 3k images annotated for 11 compound expressions.

The **EmotioNet** dataset [11] contains around 1M images and was released for the EmotioNet Challenge in 2017. 950K images were automatically annotated and the remaining 50K images were manually annotated with 11 AUs; around half of the latter constituted the validation and the other half the test set of the Challenge. Additionally, a subset of about 2.5K images was annotated with the 6 basic and 10 compound emotions.

The **DISFA** dataset [42] is a lab controlled dataset consisting of 27 videos each of which has 4,845 frames, where each frame is coded with the AU intensity on a six-point discrete scale. AU intensities equal or greater than 2 are considered as occurrence, while others are treated as non-occurrence. There are in total 12 AUs. Following usual practice, we employ subject-exclusive 3-fold cross-validation.

The **BP4D-Spontaneous** dataset[66] (in the rest of the paper we refer to it as BP4D) is a manually annotated database of spontaneous behavior containing videos of 41 subjects. There are 8 activities designed to elicit various spontaneous emotional responses, resulting in 328 video clips. A total of 140,000 frames are manually annotated. Following standard practice, we split all annotated frames into three subject-exclusive folds for 12 AUs.

The **GFT** dataset [12] consists of 96 videos of 96 subjects totalling around 130K frames. It is annotated for the occurrence of 10 AUs (1,2,4,6,10,12,14,15,23,24). The training set consists of 78 subjects of around 108K frames and the test set of 18 subjects of around 24.5K frames.

The **BP4D+** dataset [68] is an extension of BP4D incorporating more subjects (140). For each subject, 20 seconds from 4 activities are manually annotated by FACS annotators, resulting in 192,000 labelled frames. Following standard practice, we divide the subjects into four folds and use 12 AUs.

The **Aff-Wild2** dataset [26, 27, 17, 21, 18, 24, 19, 62, 16, 29, 20, 30, 32, 23, 33, 22, 31] is the largest in-the-wild dataset and the only one to be annotated in a per-frame basis for the seven basic expressions, twelve action units (AUs 1,2,4,6,7,10,12,15,23,24,25, 26) and valence and arousal (in the range $[-1, 1]$). In total, it consists of 564 videos of around 2.8M frames with 554 subjects. In this work, we utilize the set of images annotated with the 7 basic expressions, which we denote as Aff-Wild2-Expr.

The **WIDER FACE** dataset [57] comprises around 33K images sourced from the internet from a variety of real-world scenarios, featuring a total of 395K annotated faces. The annotations include bounding boxes for each face. The WiderFace dataset is diverse, covering a broad range of face variations, including different scales, poses, occlusions, expressions, and lighting conditions. The dataset is split into training (40%), validation (10%) and testing (50%) subsets by randomly sampling from 61 scene categories. The dataset is divided into three subsets —Easy, Medium, and Hard—based on the difficulty level of detecting the faces. The Easy subset includes faces that are relatively large, well-lit, and unobstructed, while the Medium subset presents more challenging conditions, such as smaller faces or mild occlusions. The Hard subset is particularly demanding, with faces that are extremely small, heavily occluded, or captured under poor lighting conditions.

The **UFDD** dataset [44] comprises 6.5K images, with a total of 11K annotated faces. The annotations are precise, with bounding boxes provided for each face. The images in UFDD are selected to represent a wide variety of challenging scenarios, including severe occlusions, extreme poses, exaggerated facial expressions, varying image quality, and challenging lighting conditions (including both overexposure and underexposure). Moreover, UFDD includes images with faces captured in different environments and contexts, such as crowded scenes, group photos, and images with background clutter, which are often missing in other datasets.

D Performance Metrics

Here we provide an overview and more detailed explanation on the various metrics that we utilize to measure both the overall performance and the fairness of behaviour analysis performed by FaceMultiBehaviourNet (component of our Behaviour4All toolkit), the various sota methods and the other existing toolkits (OpenFace, LibreFace, py-feat and AFFDEX 2.0). We also present and explain the performance metrics used to measure the face detection and landmark localization performance of FaceLocalizationNet, the sota methods and the other toolkits.

Note: We report the F1 for Basic Expression Recognition, as it is a stricter and more informative metric (than total accuracy) for imbalanced datasets like RAF-DB. In fact, on RAF-DB, high accuracy can be achieved by simply predicting the most frequent classes, e.g., a model that perfectly classifies only the 3 most frequent (and easiest) classes (neutral, happy, sad) already reaches 76.5% accuracy; if it also perfectly classifies the 4th most frequent class (surprise), accuracy rises to 87.1%.

D.1 Basic Expression Recognition (BER)

F_1 Score: It evaluates the overall expression recognition performance across the entire partition (training, validation, or test) by averaging the F_1 scores of each expression class $c \in C$. It ranges from $[0, 1]$, with higher values being more desirable. It is defined as:

$$F_1 = \frac{1}{|C|} \sum_c F_1^c \tag{15}$$

where: F_1^c is the F_1 Score for the c-th expression class across the entire partition.

Equality of Opportunity (EOP): It evaluates the fairness of basic expression recognition with respect to a demographic attribute (age, race, gender). It ensures that the model performs equally well across different subgroups of a demographic attribute. This prevents the model from favouring one subgroup over another, thus promoting fairness. It ensures that one can quantify and reduce disparities in the model's performance across different subgroups. It takes values in $[0, 1]$, with 0 indicating perfect equality of opportunity; values equal or below 0.1 are considered fair.

For each subgroup g of a demographic attribute G, at first we compute the confusion matrix C_g. We then normalize each of these matrices so that the sum of each row is 1, turning them into error rate matrices. This normalization gives the error rate of predicting class j when the true class is i. A normalised confusion matrix is defined as:

$$\hat{C}_g[i, j] = \frac{C_g[i, j]}{\sum_j C_g[i, j]}$$

Then, we compute the pairwise distances between the normalized confusion matrices using the Mean Absolute Deviation (MAD). For matrices $\hat{C}_g$ and $\hat{C}_{g'}$ (of subgroups g and g'):

$$\text{MAD}(\hat{C}_g, \hat{C}_{g'}) = \frac{\sum_i \sum_j \left| \hat{C}_g[i, j] - \hat{C}_{g'}[i, j] \right|}{N^2}$$

Finally, we aggregate the pairwise distances to get a single measure of equality of opportunity across all subgroups. We use the mean of these distances. Therefore, EOP is defined as:

$$\text{EOP} = \frac{2}{|G| \cdot (|G| - 1)} \sum_{g \in G} \sum_{g' \in G, g \neq g'} \text{MAD}(\hat{C}_g, \hat{C}_{g'}). \tag{16}$$

D.2 AU Detection (AUD)

F_1 Score: It evaluates the overall performance of AU Detection across the entire partition (i.e., training, validation, or test set). It is calculated by averaging the F_1 scores of each AU over the whole partition. It takes values in $[0, 1]$, with higher values being desirable.

Equal Opportunity Difference (EOD): It evaluates the fairness of AU detection with respect to a demographic attribute (age, race, gender). It measures fairness by ensuring that the True Positive Rates (TPR) for each AU are similar across different demographic subgroups. This helps to avoid biases where some subgroups might have significantly higher or lower TPRs for certain AUs, leading to unfair advantages or disadvantages. It takes values in $[0, 1]$; values ≤ 0.1 are considered fair.

For each subgroup $g \in G$ and each au, at first, we compute the:

$$\text{TPR}_g^{\text{au}} = \frac{\sum_{i \in G_k} \mathbb{1}\left(\hat{y}_i^{au} = 1 \wedge y_i^{au} = 1\right)}{\sum_{i \in G_k} \mathbb{1}\left(y_i^{au} = 1\right)}$$

where: G_k is the set of indices of samples belonging to subgroup g; $\hat{y}_i^{au}$ is the prediction for the specific au for sample i; y_i^{au} is the true label for the specific au for sample i.

Then, for each au, we calculate the difference between the maximum and minimum TPRs across all subgroups; finally, EOD is the average of these differences across all M AUs:

$$\text{EOD} = \frac{1}{M} \sum_{au} \left(\max_{g \in G} \text{TPR}_g^{au} - \min_{g \in G} \text{TPR}_g^{au} \right). \tag{17}$$

D.3 Valence-Arousal Estimation (VA-E)

CCC: It evaluates the overall performance of VA Estimation across the entire partition (i.e., training, validation, or test set). It takes values in $[-1, 1]$, with higher values being desirable. It is calculated:

$$\text{CCC} = \frac{1}{2} \left(\text{CCC}_{\text{V}} + \text{CCC}_{\text{A}} \right) \tag{18}$$

where: $\text{CCC}_i = 2 \cdot s_{xy} / \left[s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2 \right]$, with $i =$ V or A; s_x and s_y represent the variances of the valence/arousal annotations and predicted values, respectively; $\bar{x}$ and $\bar{y}$ denote the mean values of the corresponding annotations and predictions; s_{xy} represents their covariance.

fairness CCC (fCCC) : It evaluates the fairness of VA Estimation with respect to a demographic attribute (age, race, gender). It evaluates the alignment between predicted values and annotations across diverse demographic subgroups within a demographic attribute. It is computed by averaging the mean CCC values across all subgroups:

$$\text{fCCC} = \frac{1}{2 \cdot |G|} \sum_{g} \left(1 - \text{CCC}_{\text{V}}^{g} + 1 - \text{CCC}_{\text{A}}^{g} \right) \tag{19}$$

where: CCC_i^g denotes the CCC of valence/arousal calculated over all images of the g-th subgroup.

D.4 Face Detection

The evaluation of face detection models on WiderFace is performed using the Average Precision and mean Average Precision (mAP) metric, which is derived from the Precision-Recall (PR) curve. For each of the three difficulty settings (Easy, Medium, and Hard), at first, Precision (P) and Recall (R) are calculated at various confidence thresholds, then the PR curve is plotted by varying the detection confidence threshold, plotting Precision against Recall; finally, the Average Precision is calculated as the area under the PR curve (AUC). This is done for each difficulty setting, resulting in three AP values corresponding to the Easy, Medium, and Hard settings; their average creates the mAP.

E Training Implementation Details

To address the presence of approximately 20% tiny faces in the WIDER FACE training set, we adopt the approach outlined in previous works and randomly crop square patches from the original images and resize them to a resolution of 640×640 pixels during training. In addition to random cropping, we augment the training dataset by applying random horizontal flipping and photometric color distortion. FaceLocalizationNet is trained using SGD optimizer with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 32 across four NVIDIA Tesla P40 GPUs, each with 24GB of memory. The learning rate is initially set to 10^{-3} and is increased to 10^{-2}. The training process is concluded after 80 epochs.

F Experimental Results

F.1 Face Detection

Table 8 provides a comprehensive performance comparison (in terms of AP) for Face Detection on WIDER FACE. The comparison includes: i) the FaceLocalizationNet component of our proposed toolkit; we show 2 cases: when the backbone network is a MobileNet-0.25 (denoted as FaceLocalizationNet (M-0.25)); and when the backbone network is a ResNet50 (denoted as FaceLocalizationNet (R50)); ii) the sota methods: MTCNN [64], ScaleFace [58], CMS-RCNN [73], MSCNN [4], HR [15], SSH [45], Face R-CNN [52] and FAN [52]; and iii) online RetinaFace implementations either in Tensorflow (T) or Pytorch (P) (peteryuX-P-R50, biubug6-P-R50, biubug6-P-M0.25). It is evident that FaceLocalizationNet-Q (R50) and FaceLocalizationNet (R50) outperform all other sota methods, as well as their online implementations. Finally, let us note that the results that we provide for our models are performed in single scale, whereas most sota methods employ flipping as well as multi-scale strategies and box voting.

Table 8: Evaluation of Face Detection (in terms of AP in %) on WIDER FACE dataset by FaceLocalizationNet and the state-of-the-art methods and other implementations

Method	WIDER FACE		
	Easy	Medium	Hard
MTCNN	84.8	82.5	59.8
ScaleFace	86.8	86.7	77.2
CMS-RCNN	89.9	87.4	62.4
MSCNN	91.6	90.3	80.2
HR	92.5	91.0	80.6
SSH	93.1	92.1	84.5
Face R-CNN	93.7	92.1	83.1
SFD	93.7	92.5	85.9
Face R-FCN	94.7	93.5	87.4
FAN	95.2	94.0	84.1
peteryuX-P-R50	95.5	94.0	84.4
biubug6-P-R50	95.5	94.0	84.4
biubug6-P-M0.25	88.7	87.1	81.0
FaceLocalizationNet (M-0.25)	**93.5**	**92.0**	**79.8**
FaceLocalizationNet (R50)	**95.6**	**94.6**	**88.5**

F.2 Generalizability for Face Detection

Table 9 provides a comprehensive performance comparison for Face Detection on UFDD (in terms of mAP). The comparison includes the FaceLocalizationNet component of our proposed toolkit (the four cases that we described previously) trained on the WIDER FACE and the sota methods. Although all four versions of our model have not been trained on UFDD, they were able to surpass all sota methods that have been trained on UFDD.

Table 9: Evaluation of Face Detection (in terms of mAP in %) on UFDD dataset by FaceLocalizationNet and the state-of-the-art methods

Method	UFDD
	mAP
Faster-RCNN	52.1
SSH	695
S3FD	72.5
HR-ER	74.2
FaceLocalizationNet (M-0.25)	**76.9**
FaceLocalizationNet (R50)	**92.1**

F.3 Behavior Analysis

F.3.1 Comparison with Student–Teacher Multitask Learning

For additional comparison, we also evaluate a Student–Teacher (Stu–Tea) knowledge distillation strategy. In this setting, one or multiple teacher networks are used to generate soft pseudo-labels for datasets that originally contain annotations for only a subset of tasks. This process converts partially annotated datasets into fully overlapping multitask datasets. A multitask student network is then trained on the pseudo-completed annotations. For fair comparison, we use the same backbone architecture as in our proposed framework. Table 10 presents the comparison between: (i) single-task learning (ST), (ii) vanilla multitask learning (MT), (iii) the proposed dependency-aware framework, and (iv) the Student–Teacher multitask baseline. The Student–Teacher approach slightly improves over vanilla MTL on most datasets and tasks, indicating that pseudo-label completion partially alleviates the heterogeneous supervision problem. However, negative transfer still persists for expression recognition, where the multitask performance remains below single-task learning on AffectNet-BE and RAF-DB. In contrast, the proposed dependency-aware framework substantially outperforms the Student–Teacher baseline across all datasets and tasks. These results suggest that simply generating overlapping annotations is insufficient for stable multitask optimisation. Explicit modelling of task dependencies and representation geometry remains necessary to effectively mitigate negative transfer and conflicting gradients.

Table 10: Comparison with Student–Teacher multitask learning.

Dataset	Metric	ST	MT	Ours	Stu–Tea
AffectNet	VA	58.1	59.2	**63.7**	59.8
AffectNet	F1	64.3	62.1	**68.1**	62.6
RAF-DB	F1	68.4	65.1	**79.8**	65.6
EmotioNet	F1	51.9	53.1	**81.8**	53.5
DISFA	F1	59.9	61.1	**82.1**	61.5
BP4D	F1	79.9	81.1	**85.0**	81.5
BP4D+	F1	58.1	59.6	**75.2**	60.0
Aff-Wild	CCC	53.3	55.8	**59.3**	56.4

F.3.2 Backbone Analysis & Model Agnostic Losses

Table 11 presents the results of widely used backbones and ours. Each architecture is evaluated under 3 scenarios: i) Uni-task training (ST); ii) Multi-Task training (MT); and iii) Enhanced Multi-Task training (where the proposed losses are used - MTwL). It is evident that negative transfer occurs in the case of BER, as all uni-task models produce better performance than their MT counterparts. However, when the proposed losses are added, this is alleviated across all different backbones and their performance is significantly increased across all tasks and databases, verifying that our losses are (backbone) model agnostic (bringing similar level of improvements to all backbone networks and alleviating negative transfer in all of them). Finally, our backbone architecture has the best performance against ResNet-50 and ViT.

Table 11: Performance comparison across datasets using widely adopted backbones and our proposed model. Each architecture is evaluated under three training scenarios: single-task (ST), multi-task (MT), and multi-task with losses (MTwL).

Dataset	**Metric**	**Ours**			**ResNet-50**			**ViT**		
		ST	**MT**	**MTwL**	**ST**	**MT**	**MTwL**	**ST**	**MT**	**MTwL**
AffectNet	CCC-VA	58.1	59.2	**63.7**	57.1	58.5	62.5	57.5	58.1	62.2
	F1	64.3	62.1	**68.1**	62.9	60.1	67.8	63.3	59.2	67.8
RAF-DB	F1	68.4	65.1	**79.8**	67.4	64.0	78.7	68.1	63.3	78.9
EmotioNet	F1	51.9	53.1	**81.8**	52.4	52.6	79.8	51.5	52.9	79.8
DISFA	F1	59.9	61.1	**82.1**	58.5	60.1	81.0	59.1	59.2	80.7
BP4D	F1	79.9	81.1	**85.0**	78.4	80.1	84.9	78.9	79.5	84.6
BP4D+	F1	58.1	59.6	**75.2**	57.0	57.9	73.8	58.1	56.7	73.4
Aff-Wild	CCC	53.3	55.8	**59.3**	52.1	54.2	57.9	53.1	53.3	57.8

F.3.3 Toolkit's sota Performance due to Merged Training Sets? Proof of our contributions

One might assume that our toolkit's sota performance is primarily due to the increased amount of training data rather than the contributions of our method's components. However, this is not the case. Table 12 clearly demonstrates that our model's design and key innovations are the primary drivers of its performance. All models of Table 12 are CNNs with same architecture as FaceMultiBehaviourNet.

Row 1 shows eight Single-Task (ST) models trained on individual datasets (e.g., one model trained on Aff-Wild; another trained on RAF-DB). Row 2 shows three Single-Task (ST) models trained on grouped datasets by task (e.g. one model trained on the union of EmotioNet, DISFA, BP4D, and BP4D+; another trained on the union of AffectNet and Aff-Wild). While the models in Row 2 are trained on more data, they do not consistently outperform those in Row 1. In fact, performance drops significantly on RAF-DB and EmotioNet. This supports prior findings [59, 63] that adding more data for the same task can be harmful due to label inconsistencies, annotation errors, and dataset biases.

Row 3 further shows that more data alone does not help - training a MT model on all datasets without our proposed losses results in Negative Transfer (NT) for BER (where performance declines compared to ST models). The NT effect occurs because the VA and AU tasks, which have significantly more labeled data, dominate the learning process, effectively suppressing the expression task. As shown in prior work [55], NT is strongly influenced by task-specific data distribution, where tasks with fewer annotations struggle to contribute effectively to shared representations. Row 4 shows our proposed losses effectively mitigate NT and improve overall performance, which emphasizes the critical role of these components in stabilizing multi-task learning and enhancing overall performance.

Table 12: Performance evaluation (in %) of Behaviour4All vs. MT and ST configurations.

ID	Method	Aff-Wild	AffectNet		RAF-DB	EmotioNet	DISFA	BP4D	BP4D+
			VA	BE					
1	$8 \times$ ST	52.1	57.1	64.3	69.7	53.4	58.5	78.4	56.5
2	$3 \times$ ST	53.3	58.1	64.3	68.4	51.9	59.9	79.9	58.1
3	MT	55.8	59.2	62.1	65.1	53.1	61.1	81.1	59.6
4	**Ours**	**59.3**	**63.7**	**68.1**	**79.8**	**81.8**	**82.1**	**85.0**	**75.2**

F.3.4 Fairness

BER: Table 13 presents a detailed fairness comparison for BER across various demographics (age, race, and gender) for BER on AffectNet and RAF-DB (it also shows the overall performance shown in main paper). The results indicate that our toolkit consistently demonstrates greater fairness across all demographics and datasets when compared to both sota and the other toolkits. Notably, on both datasets utilized, the following observations can be made. For *gender*, our toolkit is unbiased, with an EOP below 10%, whereas the sota and other toolkits exhibit bias. For *race*, although our toolkit is not entirely fair, the EOP score of $<$15% suggests it is close to meeting fairness criteria, in contrast to sota and other toolkits, which are biased with EOP scores of $\geq$20%. For *age*, our toolkit does not meet fairness criteria, as the EOP scores exceed 26%.

Table 13: BER: Performance (↑) & Fairness (↓) evaluation (F1, EOP in %)

Datasets	AffectNet					RAF-DB				
	Original	New	Age	Gender	Race	Original	New	Age	Gender	Race
DAN	65.7	60.0	31.6	15.4	22.1	64.5	70.9	32.1	14.1	21.1
EAC	65.3	60.3	31.9	12.5	20.2	63.6	75.5	32.5	13.1	21.6
MA-Net	64.5	55.4	32.5	13.5	18.9	60.7	65.2	33.6	14.5	22.3
EffFace	63.7	58.6	32.1	13.1	19.5	68.7	73.2	33.4	14.0	21.5
POSTER++	67.4	59.6	31.8	12.3	19.7	68.9	73.7	32.9	13.3	22.4
UA-FER	66.9	58.9	31.9	12.3	20.2	68.8	72.3	33.3	13.5	22.1
MHAN	67.7	59.2	31.6	12.1	19.6	67.7	72.5	32.9	13.4	21.8
CLIPER	66.2	58.1	31.9	15.9	23.1	67.9	72.1	32.8	14.8	22.3
CLIP	64.7	55.9	32.8	13.6	19.3	66.6	71.5	33.7	14.8	22.7
Norface	**68.6**	60.2	32.9	11.5	19.1	69.1	74.5	33.9	12.7	21.9
py-feat	55.0	-	33.4	13.3	21.1	49.3	-	28.3	13.8	20.9
LibreFace	49.7	-	35.4	13.4	22.6	59.6	-	30.1	14.2	22.4
Ours	<u>68.1</u>	**62.9**	**27.5**	**9.2**	**14.6**	**79.8**	**80.3**	**25.5**	**8.6**	**14.1**

VA-E Table 14 presents fairness comparison for VA-E across various demographics. Results indicate that our toolkit consistently demonstrates greater fairness across all demographics compared to sota.

Table 14: VA-E: CCC Performance (↑) & fCCC Fairness (↓)

Datasets	Aff-Wild	AffectNet				
	Original	Original	New	Age	Gender	Race
FUXI	52.0	56.5	74.0	56.7	44.0	52.1
SITU	53.1	57.5	71.1	53.1	43.4	48.3
CTC	50.2	52.3	71.0	60.3	47.2	55.6
USTC-IAT	51.4	53.7	71.2	62.1	48.2	56.4
GRJCA	50.9	53.9	71.6	62.6	49.3	56.9
CtyunAI	50.6	53.5	71.2	62.8	49.9	57.9
VA-StarGan	50.2	54.5	72.4	57.8	46.4	53.1
MT-EffNet	51.7	57.2	71.8	54.8	43.7	49.2
Ours	**59.3**	**63.7**	**79.3**	**48.9**	**37.8**	**43.1**

Table 15: AUD: Fairness (↓) evaluation (EOD in %)

Datasets	EmotioNet			DISFA		
	Age	Gender	Race	Age	Gender	Race
FUXI	35.8	8.0	22.2	48.3	19.5	49.3
SITU	40.8	7.8	22.0	45.7	12.8	33.3
CLEF	41.6	6.8	21.8	42.8	12.5	32.3
AUFormer	39.9	6.9	19.8	48.3	19.6	40.2
FMAE-IAT	37.1	6.4	20.0	41.2	17.8	45.1
MDHR	39.3	8.4	23.2	41.5	17.2	44.8
MCM	38.1	7.2	22.4	41.2	17.1	44.1
py-feat	38.7	8.9	17.3	40.3	13.5	29.6
LibreFace	42.4	11.07	19.3	43.6	15.6	31.6
OpenFace	45.5	8.0	20.6	52.1	13.9	32.7
Ours	**31.1**	**4.8**	**13.1**	**31.5**	**8.5**	**17.4**

AUD Table 15 presents a fairness comparison across various demographics for AUD on EmotioNet and DISFA. Consistent with the results observed for BER, our toolkit demonstrates superior fairness across all demographics and datasets compared to sota and 3 toolkits. The findings for *race* and *age* align with those observed in BER analysis. The primary difference arises for *gender*, where our toolkit remains unbiased (EOD below 10%) across both datasets, while the sota and other toolkits are unbiased on EmotioNet but exhibit bias on DISFA.

Overall Fairness Results In affective computing and facial behavior analysis, bias often emerges because: (i) models overfit to dominant demographic groups; (ii) certain EXP or AUs are underrepresented in particular demographics, and (iii) task heads learn demographically correlated shortcuts rather than genuine affective cues. Our framework constrains how tasks and features interact, which suppresses those shortcuts and helps enforce demographic invariance.

F.3.5 Downstream Task: Compound Expression Recognition

Here, we extend our experiments by using the proposed toolkit as a foundation model, leveraging its rich facial behaviour representations for efficient transfer learning. We evaluate its zero-shot and few-shot performance on Compound Expression Recognition (CER) to assess adaptability with minimal or no task-specific data. These experiments demonstrate strong generalization, robustness to new tasks, and reduced reliance on large labeled datasets.

For the zero-shot learning experiment, we use our toolkit's predictions together with the rules from [9] to generate compound expression (CE) predictions. We compute a candidate score, $\mathcal{C}_{exp}$, for each compound expression: $\mathcal{C}_{exp} = \mathcal{I}_{AU} + \mathcal{F}_{exp} + \mathcal{D}_{VA}$, where:
(i) $\mathcal{F}_{exp} = p_{exp_1} + p_{exp_2}$: BE exp_1, exp_2 form CE (e.g., for *happily surprised*: exp_1 is *happy*, exp_2 is *surprise*);
(ii) $\mathcal{D}_{VA} = \begin{cases} 1, & p_V > 0 \\ 0, & \text{otherwise} \end{cases}$

(iii) $\mathcal{I}_{AU} = \frac{\sum_{\text{AU}_i} p_{\text{AU}_i|\text{exp}}}{\sum_{\text{AU}_i} p_{\text{AU}_i} \cdot p_{\text{AU}_i|\text{exp}}}$

The final prediction is CE with maximum candidate score.

Table 16 provides a comprehensive performance comparison for CER on EmotioNet (in terms of F1) and RAF-DB (in terms of average accuracy AA). Our toolkit consistently outperforms all sota by significant margins in both zero- and few-shot settings. As anticipated, the best overall performance is achieved by our toolkit under the few-shot setting.

Table 16: CER: Perf. evaluation (in %); 'AA' is average accuracy

datasets	EmotioNet	RAF-DB
	F1	AA
NTechLab [2]	25.5	-
VGG-FACE-mSVM [37]	-	31.6
DLP-CNN [37]	-	44.6
zero-shot Behaviour4All	**32.9**	**47.9**
fine-tuned Behaviour4All	**40.7**	**57.5**

F.3.6 Efficiency Analysis

We ran all experiments with Intel i9-13900K CPU, 64GB DDR5 4800MHz RAM, and Nvidia GTX1080 on Windows 11 Operating System. E-cores and hyper-threading were disabled. We randomly picked 5000 images and fed them to the existing toolkits and our toolkit and computed the average FPS, the total number of parameters (in millions) and the total number of Giga FLOPs. Additionally, we further compare the computational complexity of our toolkit with that of other Multi-Task (MT) methods that that jointly estimate VA, 7 basic expressions and AUs. It is important to note that Behaviour4All is fully multi-task, predicting VA, 17 AUs and 7 basic expressions simultaneously (naturally increasing parameters, FLOPs and reducing FPS), whereas the other toolkits typically predict only 7 expressions, or at most 7 expressions plus 8 AUs. The results can be seen on Table 17.

Table 17: Efficiency (in terms of FPS) and model size comparison (in terms of FLOPs, total number of parameters and size) between Behaviour4All and other toolkits

Toolkits	# Params (M)	GFLOPs	FPS
LibreFace	22.5	3.7	26.9
OpenFace	44.8	7.4	13.5
py-feat	49.3	8.2	12.2
FUXI	240	35	2
CtyunAI	115	5.5	15
Behaviour4All	*23.1*	*3.8*	*26.2*

F.3.7 Relationship between the Tasks

In the main paper we utilized the cognitive and psychological study of [9] and extracted the relationship between expressions and AUs. Here we will show that our proposed loss terms can also be used if a different relationship between expressions and AUs is considered; this time this relationship is not extracted from domain knowledge, but is inferred empirically from dataset annotations.

Assuming that K training data are annotated with a specific expression class, Exp_i, and L of these K data are annotated with a specific AU, AU_j (L<K), i.e., L data are annotated with both expression class Exp_i and AU AU_j, then $p(AU_j|Exp_i) = L/K$. We follow this procedure for all AUs of each class, thus creating the relationship between expressions and AUs of Table 18 (by using the Aff-Wild2 dataset that contains annotations for both tasks). In other words, we first find the co-occurrence of Exp_i and AU_j; we then divide this by the occurrence of Exp_i, creating $p(AU_j|Exp_i)$.

Tables 19, 21 and 20 compare the performance of our toolkit to the sota when the relationship between expressions and AUs is inferred empirically from dataset annotations (Aff-Wild2). The results reveal minimal differences between the performance of our model when trained with the two different expression-AU relationships (from Tables 2 -of the main paper- and Table 18 -of the supplementary-).

Table 18: Relationship of expressions & AUs from Aff-Wild2 (the weights in brackets correspond to the percentage of images annotated with the specific expression in which the AU was activated)

Expression	AUs (with weights w)
happiness	**12** (0.82), **25** (0.7), **6** (0.57), **7** (0.83), **10** (0.63)
sadness	**4** (0.53), **15** (0.42), **1** (0.31), **7** (0.13), **17** (0.1)
fear	**1** (0.52), **4** (0.4), **25** (0.85), **5** (0.38), **7** (0.57), **10** (0.57)
anger	**4** (0.65), **7** (0.45), **25** (0.4), **10** (0.33), **9** (0.15)
surprise	**1** (0.38), **2** (0.37), **25** (0.85), **26** (0.3), **5** (0.5), **7** (0.2)
disgust	**9** (0.21), **10** (0.85), **17** (0.23), **4** (0.6), **7** (0.75), **25** (0.8)

Table 19: Performance evaluation (F1 (↑) in %) of Basic Expression Recognition

Datasets	AffectNet		RAF-DB	
	Orig	New	Orig	New
DAN	65.7	60.0	64.5	70.9
EAC	65.3	60.3	63.6	75.5
MA-Net	64.5	55.4	60.7	65.2
EffFace	63.7	58.6	68.7	73.2
POSTER++	67.4	59.6	68.9	73.7
UA-FER	66.9	58.9	68.8	72.3
MHAN	67.7	59.2	67.7	72.5
CLIPER	66.2	58.1	67.9	72.1
CLIP	64.7	55.9	66.6	71.5
Norface	**68.6**	60.2	69.1	74.5
py-feat	55.0	-	49.3	-
LibreFace	49.7	-	59.6	-
Ours	<u>68.3</u>	**63.2**	**79.9**	**80.0**

Table 20: Performance evaluation (CCC (↑) in %) of VA Estimation

Datasets	Aff-Wild	AffectNet	
	Original	Original	New
FUXI	52.0	56.5	74.0
SITU	53.1	57.5	71.1
CTC	50.2	52.3	71.0
USTC-IAT	51.4	53.7	71.2
GRJCA	50.9	53.9	71.6
CtyunAI	50.6	53.5	71.2
VA-StarGan	50.2	54.5	72.4
MT-EffNet	51.7	57.2	71.8
Ours	**59.1**	**63.5**	**79.5**

Table 21: AUD: Performance (↑) evaluation (F1 in %)

Datasets	EmotioNet		DISFA		BP4D	BP4D+
	Original	New	Original	New	Original	Original
FUXI	60.2	77.9	49.8	43.7	63.2	56.2
SITU	59.7	77.2	49.2	40.4	62.7	55.8
CLEF	57.6	74.6	64.8	45.9	65.9	63.1
AUFormer	63.6	78.8	66.4	46.2	66.2	62.7
FMAE-IAT	59.8	72.9	70.1	46.1	67.1	66.8
MDHR	60.6	77.2	64.9	45.1	66.6	63.4
MCM	58.6	76.4	64.3	43.8	66.0	62.7
py-feat	36.5	-	54.0	-	61.4	52.4
LibreFace	35.7	-	63.0	-	62.0	55.6
OpenFace	35.5	-	59.0	-	61.3	53.5
Ours	**81.5**	**87.8**	**81.9**	**68.5**	**84.7**	**75.5**

F.3.8 Ablation on Losses

In Table 22, we conduct a comprehensive ablation study to assess the impact of varying the λ values, which control the contribution of each proposed loss term on the toolkit's performance across multiple datasets and tasks. It can be seen that each loss terms brings a contribution to each of the 3 behavior tasks (VA-E, AU-D and EXP recognition). Overall, the full model consistently achieves the best results (while alleviating negative transfer).

Table 22: Ablation study for BER, AUD and VA-E on proposed losses; CCC is performance metric for Aff-Wild, AffectNet-VA; F1 score is performance metric for the rest datasets; Aff-W stands for Aff-Wild; Aff-VA and Aff-BE stand for AffectNet-VA and AffectNet-BE; Enet stands for Emotionet

$(\lambda_s, \lambda_g, \lambda_f, \lambda_r, \lambda_c, \lambda_{cw\text{-}ds}, \lambda_g)$	Aff-W	Aff-VA	Aff-BE	RAF-DB	Enet	DISFA	BP4D	BP4D+
(0, 0, 0, 0, 0, 0, 0)	55.8	59.2	62.1	65.1	53.1	61.1	81.1	59.6
(0.05, 0, 0, 0, 0, 0, 0)	56.3	60.4	63.0	67.5	54.1	62.0	82.0	61.1
(0.1, 0, 0, 0, 0, 0, 0)	56.5	60.8	63.5	68.5	54.5	62.5	82.5	61.3
(1.0, 0, 0, 0, 0, 0, 0)	56.2	60.0	61.6	64.7	53.6	61.7	81.7	60.2
(0.1, 0.05, 0, 0, 0, 0, 0)	56.5	61.1	64.4	69.9	64.5	70.5	83.5	68.1
(0.1, 0.1, 0, 0, 0, 0, 0)	56.5	60.9	63.8	68.9	58.2	66.1	82.7	64.2
(0.1, 0.05, 0.05, 0, 0, 0, 0)	56.6	61.2	65.0	70.8	65.1	71.0	83.8	68.7
(0.1, 0.05, 0.1, 0, 0, 0, 0)	56.6	61.3	65.3	71.5	65.5	71.5	84.0	69.1
(0.1, 0.05, 1,. 0, 0, 0, 0)	56.5	61.1	64.8	70.4	64.7	70.7	83.6	68.3
(0.1, 0.05, 0.1, 0.05, 0, 0, 0)	57.4	62.3	65.8	73.2	70.5	75.5	84.7	73.2
(0.1, 0.05, 0.1, 0.1, 0, 0, 0)	56.9	61.7	65.5	72.6	68.7	73.8	84.1	71.5
(0.1, 0.05, 0.1, 0.05, 0.05, 0, 0)	57.5	62.4	66.4	75.2	71.5	76.5	84.9	74.2
(0.1, 0.05, 0.1, 0.05, 0.1, 0, 0)	57.4	62.3	66.1	73.7	71.0	76.0	84.7	73.5
(0.1, 0.05, 0.1, 0.05, 0.05, 0.05, 0)	58.3	63.0	67.1	77.1	76.8	80.1	85.0	75.0
(0.1, 0.05, 0.1, 0.05, 0.05, 0.1, 0)	57.8	62.6	66.9	76.4	73.7	78.8	84.9	74.6
(0.1, 0.05, 0.1, 0.05, 0.05, 0.05, 0.1)	58.6	63.3	68.1	79.4	78.2	81.1	85.0	75.1
(0.1, 0.05, 0.1, 0.05, 0.05, 0.05, 0.05)	59.3	63.7	68.1	79.8	81.8	82.1	85.0	75.2

F.3.9 Optimisation Dynamics and Gradient Conflict Analysis

Behaviour of Losses During Training The optimisation behaviour of Behaviour4All is structured by design through a three-stage optimisation strategy.

Stage 1: Task-only warm-up (epochs 1–6). During the initial stage, only the task-specific losses are active. This stage allows the network to first learn strong task-specific representations without cross-task constraints.

Stage 2: Prior-based coupling (epochs 7–27). In the second stage, the cross-task dependency losses are progressively introduced and gradually ramped up. During this phase, the task losses continue decreasing, while the cross-task losses initially increase slightly due to the introduction of consistency constraints across tasks. As optimisation progresses and predictions become increasingly aligned across tasks, the cross-task losses subsequently decrease.

Stage 3: Geometry shaping (epochs 28–50). In the final stage, all components operate jointly, including the geometry modelling objectives. At this point, all losses decrease simultaneously and smoothly, indicating stable optimisation and coherent multitask learning.

The staged optimisation strategy plays an important role in stabilising training. Rather than enforcing all constraints simultaneously from the beginning, the framework progressively introduces relational and geometric constraints after the model has already learned meaningful task-specific structure. This improves convergence stability, reduces optimisation instability and facilitates effective interaction between tasks.

Mitigation of Gradient Conflicts To directly assess gradient conflict, we perform an explicit gradient-level analysis on the shared backbone parameters.

For each mini-batch, we compute task-specific gradients with respect to the shared backbone for the three primary objectives: (i) valence–arousal regression using CCC loss, (ii) expression recognition using cross-entropy loss, and (iii) AU detection using binary cross-entropy loss.

Denoting these gradients as: g_{VA}, g_{EXP}, g_{AU}, we measure pairwise cosine similarity between all task pairs: $\cos(g_{\text{VA}}, g_{\text{EXP}})$, $\cos(g_{\text{VA}}, g_{\text{AU}})$, $\cos(g_{\text{EXP}}, g_{\text{AU}})$.

We additionally report the *conflict rate*, defined as the percentage of mini-batches where the cosine similarity is negative.

Vanilla MTL. Under vanilla multitask learning, gradient interactions exhibit substantial conflict. The average cosine similarity between task gradients tends to become negative during training, typically ranging from approximately -0.15 to -0.25, while the conflict rate exceeds 45%. This indicates frequent opposing or weakly aligned gradient directions between tasks.

This behaviour is consistent with the observed optimisation dynamics of vanilla MTL, where task losses decrease unevenly and negative transfer occurs.

Behaviour4All. In contrast, our framework substantially improves gradient alignment throughout training. During the early optimisation stage, cosine similarities remain close to zero, typically ranging from approximately -0.05 to 0.10. As the relational and geometry components are progressively introduced, gradient alignment improves steadily, with cosine similarities increasing to positive values in the range of approximately 0.25–0.40, while the conflict rate drops below 15%. During the final optimisation stage, gradients become strongly aligned, with average cosine similarity stabilising around 0.45–0.60 and conflict rate decreasing below 5%. This behaviour is consistent with the staged loss dynamics described previously. As optimisation progresses, task gradients become increasingly compatible, resulting in simultaneous and stable decrease of all objectives. The evolution of the task losses also becomes smoother and more coherent compared with vanilla MTL. Overall, the progressive increase in gradient alignment correlates strongly with the elimination of negative transfer and the stabilisation of joint optimisation. These results provide direct evidence that the proposed dependency-aware optimisation framework effectively mitigates gradient conflict and aligns task objectives in practice.

G Generality Beyond Affective Computing

The contribution of Behaviour4All is not limited to a facial behaviour analysis toolkit. At its core, our method is a general MTL framework that reformulates task analysis as a dependency-aware optimisation problem. The proposed framework is designed to address key limitations of standard MTL, including negative transfer, conflicting gradients, learning under partially overlapping annotations and poor generalisation across datasets. To demonstrate that the proposed formulation is not specific to affective computing, we conduct an additional case study on three non-affective domains: (i) face recognition with identity and attributes using CelebA [39], (ii) shoe type recognition with class and attributes using SADD [35], and (iii) clothing recognition with category and attributes using the Clothing Attributes Dataset (CAD) [6]. In these experiments, each problem consists of two tasks: class recognition and attribute detection. The task-specific losses are defined analogously to the EXP and AU losses in the main paper. Task relationships are inferred empirically from dataset annotations. This follows the same principle as the empirical dependency estimation used for task relationships in the main framework.

G.1 Datasets and Protocol

CelebA CelebA contains face images annotated with identity and facial attributes. The original split is subject-independent. For this case study, we additionally create three subject-dependent splits.

SADD The Shoes Attribute Discovery Dataset (SADD) contains shoe images with type and attribute annotations. Since no predefined split is provided, we use 7.4K images for training and 7.3K images for testing.

CAD The Clothing Attributes Dataset (CAD) contains clothing images annotated with categories and semantic attributes. Due to its small size and lack of a predefined split, we use 6×2-fold cross-validation with 50–50 train–test splits.

We evaluate all methods using F1 score. We compare three training strategies: (i) two independent single-task models ($2 \times$ ST), (ii) vanilla MTL with a shared backbone and task-specific heads, and (iii) our dependency-aware MTL framework. We use ResNet-50 (RES) and DenseNet-121 (DEN) as backbones.

Table 23: Generality study beyond affective computing. Results are reported using F1 score. For CAD, results are shown as mean ± spread.

Model	Dataset	2 × ST: Class / Attr	MTL: Class / Attr	Ours: Class / Attr
RES	CelebA	72.9 / 71.4	75.1 / 72.3	**85.1 / 77.9**
RES	SADD	74.6 / 90.6	75.1 / 90.1	**83.4 / 96.5**
RES	CAD	41±7 / 44±2	35±9 / 38±4	**60±5 / 53±1**
DEN	CelebA	72.2 / 70.2	74.1 / 71.0	**84.0 / 76.6**
DEN	SADD	73.1 / 89.6	73.7 / 88.9	**82.0 / 95.7**
DEN	CAD	39±7 / 42±2	33±9 / 36±4	**58±5 / 51±1**

The results in Table 23 show that the proposed framework generalises beyond affective computing. On CelebA, vanilla MTL does not suffer from negative transfer and slightly improves over single-task learning. Nevertheless, our framework substantially outperforms both single-task and vanilla MTL models on both class and attribute prediction.

On SADD, vanilla MTL produces mild negative transfer for attribute detection, where attribute performance decreases compared with single-task training. Our framework alleviates this negative transfer and improves both shoe type recognition and attribute detection.

On CAD, vanilla MTL performs significantly worse than single-task learning on both tasks, indicating severe negative transfer. This is likely due to the small dataset size and severe sparsity in attribute annotations, with 12%–84% missing attribute labels across attributes. Our framework alleviates negative transfer, significantly improves over both single-task and vanilla MTL baselines, and reduces the performance spread across folds.

We demonstrate that the proposed dependency-aware MTL formulation is not tied to affective computing. Instead, it provides a general framework for learning under structured task dependencies and partially overlapping annotations.

Impact Statement

Our proposed toolkit, Behaviour4All, offers a unified framework for multi-task facial behaviour analysis in the wild, with potential positive societal implications across multiple domains. Applications such as affective computing, mental health research, education, and human-computer interaction may benefit from improved, generalisable models for analysing facial behaviour. By supporting valence–arousal estimation, expression recognition, and action unit detection, our toolkit can serve as a visual component of broader affect-understanding systems. However, facial behaviour alone does not fully capture an individual's emotional state, which may also depend on audio, language, situational context, and individual differences. Therefore, Behaviour4All should not be treated as a standalone solution for interpreting emotions or making high-stakes decisions.

However, we also acknowledge potential negative societal impacts. Facial behaviour analysis systems, when misused, can lead to privacy violations, unconsented surveillance, profiling, and misinterpretation of affective states, particularly in sensitive or high-stakes environments (e.g., law enforcement, hiring). Furthermore, even though we train on diverse, in-the-wild datasets, demographic biases inherent in public datasets may persist, potentially resulting in unequal performance across demographics.

To mitigate these concerns: (i) we use only datasets acquired with formal consent protocols (e.g., EULAs); (ii) we will release Behaviour4All with a custom End User License Agreement (EULA) that explicitly restricts usage in ethically sensitive or harmful scenarios such as surveillance, social scoring, or discriminatory applications; (iii) we will include documentation highlighting the toolkit's intended scope, known limitations, and guidelines for responsible use, and we encourage downstream developers to conduct fairness and bias evaluations tailored to their target applications.

Limitations

Although Behaviour4All achieves strong performance across multiple datasets and tasks, a few limitations remain. First, our toolkit demonstrates unbiased performance for gender, with fairness

scores (EOP or fCCC/EOD) consistently below 10% across tasks. For race, while not fully fair, our model stays close to the accepted fairness threshold with scores around 15%. However, for age our toolkit is biased, as scores exceed 26%, indicating persistent performance gaps across age groups; such performance disparities persist due to underlying dataset biases. Second, demographic fairness is evaluated only on datasets where demographic annotations are available, limiting fairness analysis for some benchmarks. Finally, Behaviour4All is currently focused on visual facial behaviour and does not yet incorporate audio, language or contextual cues, which are important for fully understanding human affect in real-world interactions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope. The paper does not present Behaviour4All merely as a toolkit, but as a dependency-aware multitask learning framework for facial behaviour analysis. The claims regarding task-specific supervision, cross-task dependency modelling, geometry modelling, mitigation of negative transfer, learning under heterogeneous annotations, and improved generalisation are supported by the method, ablation studies, gradient conflict analysis, and experiments across multiple datasets.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included a limitations section. Although Behaviour4All achieves strong performance across multiple datasets and tasks, a few limitations remain. First, our toolkit demonstrates unbiased performance for gender, with fairness scores (EOP or fCCC/EOD) consistently below 10% across tasks. For race, while not fully fair, our model stays close to the accepted fairness threshold with scores around 15%. However, for age our toolkit is biased, as scores exceed 26%, indicating persistent performance gaps across age groups; such performance disparities persist due to underlying dataset biases. Second, demographic fairness is evaluated only on datasets where demographic annotations are available, limiting fairness analysis for some benchmarks. Finally, Behaviour4All is currently focused on visual facial behaviour and does not yet incorporate audio, language or contextual cues, which are important for fully understanding human affect in real-world interactions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not contain any formal theoretical results, assumptions, or mathematical proofs. The work is proposing a deep learning toolkit for automatic facial behaviour analysis. It introduces architectural designs, training strategies, and empirical evaluations across multiple datasets, but it does not claim or present any theoretical guarantees, derivations, or theorems that require formal proofs or assumptions.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses sufficient information to reproduce the main experimental results to the extent that it supports the key claims and conclusions, by including: detailed model architecture and loss functions; detailed training setup and implementation details; all datasets used are publicly available and listed clearly, along with their use in training and evaluation; detailed evaluation protocol; detailed ablation studies and comparison tables. Finally, let us mention that the code and pre-trained models are made publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The toolkit is made publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the key training and testing details necessary to understand and interpret the results, including the data splits, the optimizer and hyperparameters, the loss terms and their weights, the training strategy and model architecture.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not include error bars or statistical significance testing.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information about the computational resources used to reproduce the experiments. Specifically, the paper details the training hardware (for FaceLocalizationNet and FaceMultiBehaviourNet), the training configurations, the input resolutions and the inference environment (for efficiency analysis).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conducted in the paper conforms in all respects with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper contains a section in the technical appendix that discusses potential positive and negative societal impacts of the work. In more detail, our proposed toolkit, Behaviour4All, offers a unified and open-source framework for multi-task facial behaviour analysis in the wild, with potential positive societal implications across multiple domains. Applications (to name a few) such as mental health monitoring, affective computing, education, and human-computer interaction stand to benefit from improved, generalizable models capable of understanding nuanced human emotions and behaviours. By supporting tasks such as valence-arousal estimation, expression recognition, and action unit detection, our toolkit may contribute to more empathetic and adaptive AI systems. However, we also acknowledge potential negative societal impacts. Facial behaviour analysis systems, when misused, can lead to privacy violations, unconsented surveillance, profiling, and misinterpretation of affective states, particularly in sensitive or high-stakes environments (e.g., law enforcement, hiring). Furthermore, even though we train on diverse, in-the-wild datasets, demographic biases inherent in public datasets may persist, potentially resulting in unequal performance across demographics. To mitigate these concerns: (i) we use only datasets acquired with formal consent protocols (e.g., EULAs); (ii) we will release Behaviour4All with a custom End User License Agreement (EULA) that explicitly restricts usage in ethically sensitive or harmful scenarios such as surveillance, social scoring, or discriminatory applications; (iii) we will include documentation highlighting the toolkit's intended scope, known limitations, and guidelines for responsible use, and we encourage downstream developers to conduct fairness and bias evaluations tailored to their target applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We have mentioned that we include safeguards as part of the public release of the toolkit Specifically, we : (i) release the Behaviour4All toolkit with a custom End User License Agreement (EULA) that explicitly restricts unethical use cases, such as surveillance, profiling, or discrimination; (ii) provide documentation outlining the intended use cases, known limitations, and ethical usage guidelines; (iii) emphasize the importance of fairness, transparency, and community accountability in downstream use.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses several publicly available datasets, and their names are properly cited and acknowledged throughout the text. These datasets are widely used in the facial behaviour analysis community and are released under explicit usage terms or EULAs, which we note were signed and respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a new asset, the Behaviour4All toolkit, a multi-task facial behaviour analysis framework; the toolkit is released publicly with: (i) a custom EULA to govern responsible use, (ii) documentation describing intended use cases, limitations, and ethical guidelines, and (iii) code, details and evaluation protocols.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The authors did not conduct any new experiments involving human subjects or crowdsourcing. Instead, the study uses pre-existing, publicly available datasets, all of which were collected and annotated by third parties under their own consent protocols and licensing terms.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, nor a study with human participants. The paper relies exclusively on publicly available datasets that were acquired through formal consent protocols, including the signing of End User License Agreements (EULAs). These datasets were shared with the authors under the condition that appropriate consent had already been obtained by the dataset providers. While the manuscript does not explicitly detail this, all datasets used in the study were released with ethical approval or consent mechanisms in place, and no additional data collection involving human participants was conducted by the authors. Therefore, the study did not require an IRB approval beyond what was established by the original dataset curators.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.