END-TO-END FULL VIDEO-LEVEL SUPERVISION FOR FACIAL DEPRESSION ASSESSMENT MODEL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Depressive states are often reflected by long-term, variable-duration facial behaviours, yet existing video-based automatic depression assessment (ADA) approaches frequently model depression from short fixed-length segments (including frames) due to memory constraints. Segment-level supervision not only limits their ability to learn cross-segment varying-length facial behavioural patterns but also fails to capture depressive facial patterns at the full-video level. In this paper, we propose Dual-phase Memory-efficient Video-level Supervision (DMVS), a novel training strategy that equips existing segment-level face-video ADA models with end-to-end full video-level supervision through depression relevance-weighted gradient modulation. Specifically, the first phase constructs variable-length facial behaviour representations by combining features extracted from short, fixed-length segments from the entire face video, followed by estimating the depression relevance of each behaviour representation and segment-feature, based on which a video-level loss is computed from relevance-weighted behaviour representations. Then, the second phase uses each estimated depression relevance to weight the corresponding segment's contribution to the backbone gradients, guiding feature learning towards depression-informative behaviours by accumulating relevance-weighted gradients from all face video segments. This way, the target ADA model is updated based on the weighted video-level gradients. Experiments on AVEC 2013 and 2014 demonstrate that our DMVS strategy consistently improves ADA performance across ADA-specific and general-purpose spatio-temporal models, supporting our DMVS as a model-agnostic and plug-and-play training strategy. Our code is provided in Supplementary Material.

1 INTRODUCTION

Depression is one of the most prevalent mental health disorders worldwide, characterised by persistent low mood and diminished interest in daily activities (Rottenberg, 2017), which causes a wide range of negative impacts from compromised physical health to reduced financial stability (Pinto & Parente, 2024). Inspired by the fact that human facial behaviours not only are informative for indicating human depression states (Ellgring, 2007; Gehricke & Shapiro, 2000) but also can be captured unobtrusively by commodity cameras, a large number of studies have explored face video-based automatic depression assessment (ADA) to enable easily accessible, fast and objective depression monitoring (Song et al., 2022; Pan et al., 2023).

While recent advanced deep learning (DL) models enable end-to-end learning depression-related spatio-temporal facial cues for video-based ADA (He et al., 2024; Wu et al., 2025; Feng et al., 2026) compared to traditional hand-crafted competitors, they are typically limited by severe memory and computational bottlenecks for directly processing high-dimensional long face videos (He et al., 2025; De Melo et al., 2019). To tackle these issues, most previous face video-based solutions (Niu et al., 2024a; Zhang et al., 2023) extract depression-related facial cues from individual frames or fixed-length short video segments, with some of them further conducting video-level aggregation on these frame/short segment-level facial features (Song et al., 2022; Xu et al., 2024; Song et al., 2018). However, these solutions fail to directly model long-term facial patterns reflected by the entire video at the initial feature extraction stage, leading to limited ability to capture essential long-range temporal dependencies and subtle facial dynamics. Specifically, these solutions are frequently achieved by first conducting **frame/segment-level depression modelling** which splits the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

face video into a set of sampled frames/fixed-length segments, where a network (e.g., 2D/3D CNN or Transformer) encodes each frame/segment as a frame/segment-level prediction (Uddin et al., 2020; He et al., 2021) or representation (Xu et al., 2024; Niu et al., 2020; De Melo et al., 2020b). Then, **video-level feature aggregation / prediction** is achieved by aggregating all frame/segment-level features or predictions of the target video into a single video-level representation/prediction via average/max pooling (De Melo et al., 2020b), attention-weighted pooling (Niu et al., 2020) or spectral representation (Song et al., 2018; Xu et al., 2024), from which the final video-level depression severity is predicted.

However, depressive facial behaviours usually unfold over varying temporal durations from fleeting expressions to sustained affective states (Jeganathan et al., 2022; Girard et al., 2014). As a result, ADA methods developed on fixed-length face video segments usually experience three main limitations: (i) non-overlapping fixed-length segments can cause long-term depression-related facial behaviours to be fragmented across multiple segments, which disrupts their semantic coherence, thereby hindering the model's ability to capture meaningful depression-related behaviour-level cues (De la Torre et al., 2007; Niu et al., 2021; De Melo et al., 2020a); (ii) brief depression-informative facial behaviours tend to be under-represented when embedded in relatively long but fixed-length segments containing predominantly non-informative expressions (Liu et al., 2026); and (iii) two-stage video-level aggregation on fixed frame/segment-level representations limits its capacity in modelling depression-related facial behavioural patterns across segments (Liao et al., 2026; Xu et al., 2024). On the other hand, due to memory and computational constraints, current end-to-end (one stage) video-level ADA modelling methods (Wu et al., 2025; de Melo et al., 2021) typically rely on aggressive video temporal down-sampling (e.g., using only 16 frames to represent a face video containing more than 10,000 frames (de Melo et al., 2021)), where depression-related local temporal facial-behaviour cues are inevitably lost (limitation iv).

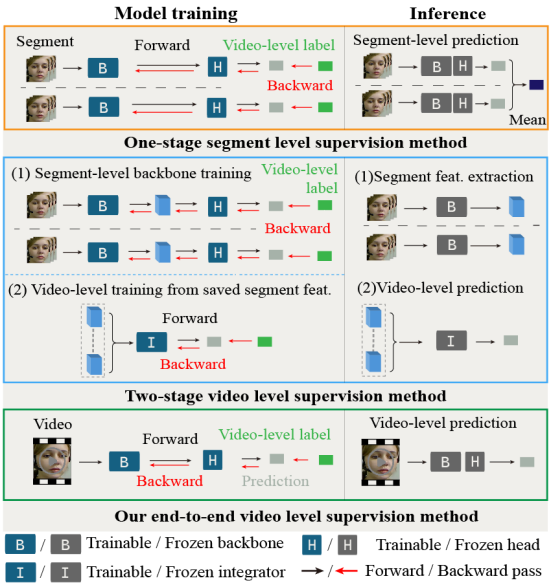


Figure 1: Comparison of training strategies for face video-based ADA. **Upper:** segment-level training with video labels, followed by averaging segment predictions. **Middle:** segment features are aggregated for video-level prediction, while the backbone remains trained on individual segments. **Bottom:** our strategy trains segment-based models end-to-end with all video segments and directly optimises video-level prediction.

In this paper, we propose Dual-phase Memory-efficient Video-level Supervision (DMVS), a novel and generic training strategy that computes the depression relevance of variable-length facial behaviours to guide the feature learning, which can enhance existing face segment-level ADA methods by introducing full video-level supervision. Specifically, the first phase groups multiple consecutive fixed-length short segments, termed as facial behaviour units (FBUs), into variable-length facial behaviour segments, aiming to capture depressive patterns at multiple temporal scales, based on which a video-level loss is computed from selected facial segment-level representations weighted by their estimated relevance with depression (**addressing limitations i and ii**). In the second phase, the backbone gradients computed from each FBU are weighted according to the depression relevance of its related facial behaviour segments. This way, our DMVS connects variable-length facial behaviours to backbone network's training under full video-level supervision (**addressing limitation iii**). We integrate activation recomputation (Chen et al., 2016) and feature-gradient caching (Gao et al., 2021) into the two-phase training process to reduce backbone activation memory, enabling end-to-end training on full videos without temporal frame subsampling (**addressing limitation iv**). Fig. 1 compares our end-to-end DMVS with existing video-based ADA model training strategies. The main contributions of this paper are summarised as follows:

- An end-to-end full-video supervision strategy for face video-based ADA that models variable-length facial behaviour segments and weights their representations by estimated depression relevance, enabling segment features to be optimised jointly within full-video behavioural context rather than independently under segment-level supervision (Sec. 3.2).
- A depression relevance-weighted gradient modulation strategy that maps segment relevance to individual facial behaviour units and weights their contributions to backbone updates under the video-level loss (Sec. 3.3).
- Extensive experiments validate DMVS as a model-agnostic and plug-and-play training strategy that reduces RMSE across five spatio-temporal models on AVEC datasets, including ADA-specific and general-purpose models (Tables 2 and 3).

2 Related Work

Frame/segment-level face-based automatic depression assessment: Due to computational constraints, most existing video-based ADA approaches analyse depression-related cues either at frame level or fixed-length segment level, which restricts their ability in capturing depression-related facial behaviours of varying temporal durations. Specifically, the backbone networks of frame-level ADA methods (Uddin et al., 2020; He et al., 2021; De Melo et al., 2019; Zhou et al., 2018) are typically trained by pairing each static facial frame with its corresponding video-level depression state, followed by video-level prediction aggregation. These backbones focus on static appearance cues associated with depression but ignore temporal dependencies among consecutive frames. Similarly, segment-level ADA methods (Zhou et al., 2020; Zhang et al., 2023; He et al., 2025; Pan et al., 2024; Niu et al., 2024a) divide videos into fixed-length facial behaviour segments, and then learn a depression-related short-term spatio-temporal facial representation from each segment and then aggregate segment-level predictions for video-level assessment, which fail to model long-term depression-related facial behaviours beyond the segment length. Alternatively, some studies (Song et al., 2022; Niu et al., 2020; Yan et al., 2024; Xu et al., 2024; He et al., 2024) further integrate all frame- or segment-level features through attention weighted pooling (Niu et al., 2020), graph representation (Xu et al., 2024; Shen et al., 2024), spectral representation (Song et al., 2022) and temporal dependency modelling (Fan et al., 2024; Uddin et al., 2022) as video-level representations for ADA. Liao et al. (2026) instead select depression-informative segments by weakly supervised scoring (Bilen & Vedaldi, 2016) from all pre-computed segment-level features. However, such two-stage pipelines operate on already fixed representations, limiting models' capacity in capturing cross-segment, variable-duration depression behavioural cues.

Video-level face-based automatic depression assessment: To capture more holistic video-level information, some ADA approaches (Song et al., 2018; de Melo et al., 2021; Wu et al., 2025) directly construct video-level representation to infer depression, although they fail to include all facial behaviours provided by the video as they remove frames/spectral components during the video-level representation learning. Specifically, de Melo et al. (2021) introduced the Maximization–Differentiation Network (MDN), where each video is represented for training by 16 consecutive frames sampled at intervals of four frames. Wu et al. (2025) first partition each video into equal-length segments, with each represented by a single frame. Then, all selected frames are directly integrated into a novel matrix graph as nodes, with matrix edges describing video-level spatio-temporal patterns, where a matrix graph neural network is designed to capture long-term depression-specific facial cues from variable-length videos.

3 Methodology

Since human depression states are usually more reliably reflected by long-term and variable-length behaviours (Krause et al., 2021; Song et al., 2022), our DMVS provides **direct video-level supervision over long and variable-length videos**, which estimates the depression relevance of every facial behaviour and subsequently applies such relevance scores to guide the backbone training. Fig. 2 illustrates its two-phase training pipeline. It should be emphasised that our DMVS is different from existing strategies (He et al., 2025; de Melo et al., 2021; Pan et al., 2024) for training video-based ADA DL models, which frequently re-utilise video-level depression labels as fixed-length short

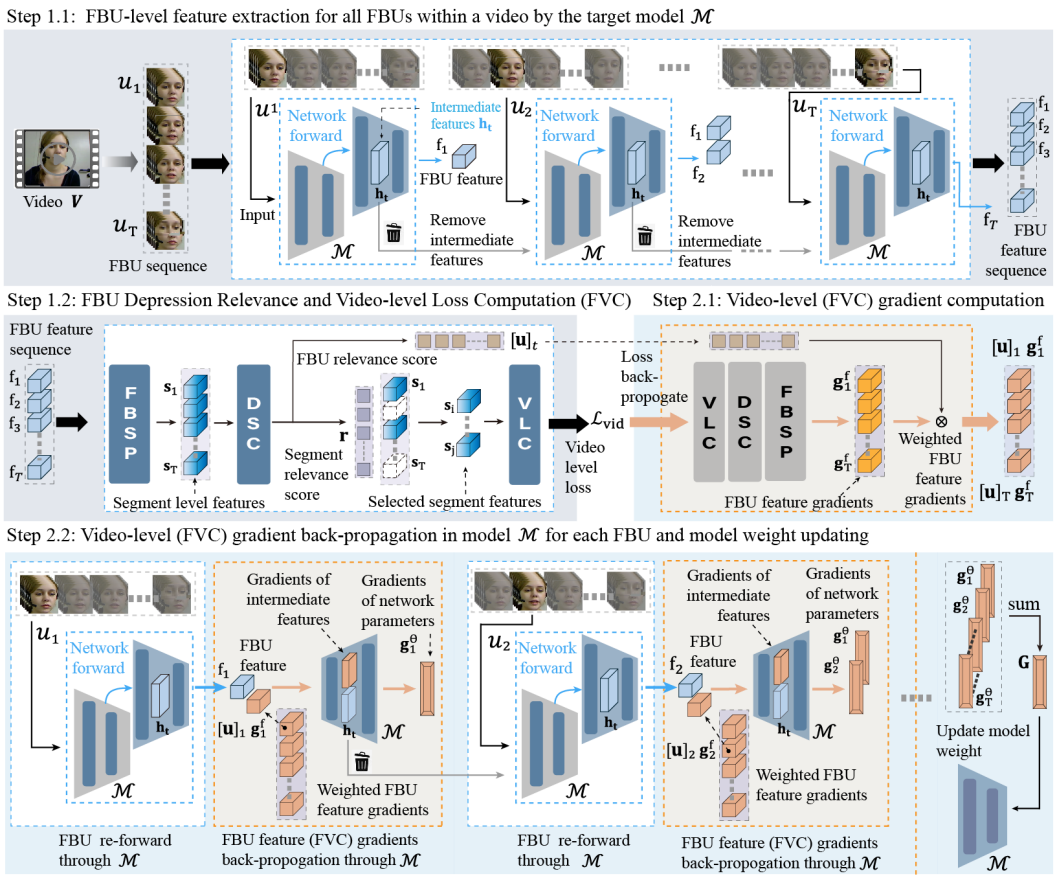


Figure 2: DMVS trains an ADA model $\mathcal{M}$ using full-video supervision and depression relevance-weighted gradient modulation. Step 1.1 sequentially extracts and caches features from all non-overlapping 32-frame facial behaviour units (FBUs), discarding backbone intermediate activations. Step 1.2 applies FVC to compute per-FBU depression relevance and a video-level loss (Fig. 3). Step 2.1 computes the gradient of this loss with respect to each cached FBU feature. Step 2.2 weights these feature gradients by per-FBU relevance and recomputes the discarded backbone activations, for back-propagation one FBU at a time. Activations are released after each FBU, and the accumulated parameter gradients update the backbone once per video. This limits backbone activation storage to one FBU.

video segment-level labels, making them mainly focus on learning depression features reflected by fixed-length short-term facial behaviours.

3.1 DMVS PIPELINE

As illustrated in Fig. 2, given a face video V that is partitioned into T consecutive *equal-length* facial behaviour units (FBUs) $\{u_t\}_{t=1}^{T}$ (each u_t is a non-overlapping 32-frame video segment), our DMVS for optimising the target ADA DL model $\mathcal{M}$ with weight θ is achieved by two phases: (i) FBU Depression Relevance and Video-level Loss Computation (FVC), and (ii) Gradient Modulation and Video-level Loss Back-propagation (GMVB). For video-level computation, we sequentially pass every FBU u_t in the video through the backbone $\mathcal{M}$ to obtain its feature $\mathbf{f}_t = \mathcal{M}(u_t; \theta)$ (Step 1.1 in Fig. 2). We cache all output features $\{\mathbf{f}_t\}_{t=1}^{T}$ but discard the backbone's intermediate activations and computation graph after processing each FBU to reduce memory use. We treat the cached features from all FBUs in the video as differentiable inputs to FVC.

The **FVC phase** predicts *variable-length* depression-related **facial behaviour segments** centred at each FBU, with each comprising multiple consecutive FBUs of a predicted duration. By considering those facial behaviour segments, the FVC computes two outputs: (1) a video-level depression

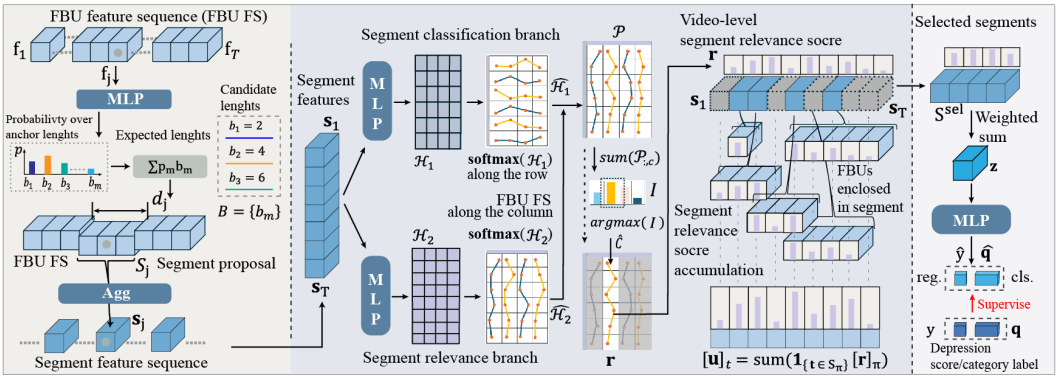


Figure 3: The FVC phase consists of three main steps: (i) **FBSP** adaptively constructs varying-length depression-related facial behaviour segments by grouping consecutive FBUs, which are then aggregated to obtain a segment representation. (ii) **DSC** selects the most depression-related segments and then evaluates the relevance of every FBU to the video-level depression status; and (iii) **VLC** computes losses based on the selected segments and per-FBU relevance score, providing full video-level supervision for optimising model $\mathcal{M}$.

prediction loss value $\mathcal{L}_{\text{vid}}$ and (2) a per-FBU relevance vector $\mathbf{u} \in \mathbb{R}^T$, where each element $[\mathbf{u}]_t$ quantifies the relevance of the corresponding FBU u_t to video-level depression status label y as:

$$\mathcal{L}_{\text{vid}}, \mathbf{u} = \text{FVC}(\text{Seg}(\{\mathcal{M}(u_t)\}_{t=1}^{T}), y). \tag{1}$$

where $\text{Seg}(\cdot)$ constructs varying-length facial behaviour segment representations from the cached full-video feature sequence $\{\mathbf{f}_t\}_{t=1}^{T}$. FVC uses these segment representations to estimate FBU depression relevance and compute the video-level loss (Step 1.2 in Fig. 2). Then, the **GMVB phase** computes the gradients of the video-level loss with respect to the FBU features, $\{\mathbf{g}_t^{\mathbf{f}} = \partial\mathcal{L}_{\text{vid}}/\partial\mathbf{f}_t\}_{t=1}^{T}$ (Step 2.1 in Fig. 2). It weights each gradient using the corresponding depression relevance score $[\mathbf{u}]_t$, enabling the estimated FBU depression relevance scores to be considered during the backbone model's updating. Since the backbone's intermediate activations were discarded during feature extraction, we recompute these activations and back-propagate through the backbone one FBU at a time. Specifically, for each FBU u_t, we recompute $\mathbf{h}_t = \mathcal{M}_{\text{int}}(u_t; \theta)$ and use the corresponding relevance-weighted feature gradient to back-propagate through the backbone (Step 2.2 in Fig. 2). This can be formulated as:

$$\theta \leftarrow \text{Update}\left(\theta, \sum_{t=1}^{T} \text{GMVB}\big(\mathbf{h}_t, [\mathbf{u}]_t\, \mathbf{g}_t^{\mathbf{f}}\big), \eta\right). \tag{2}$$

where the relevance-weighted gradients jointly update the backbone weights θ in a single video-level optimisation step with a learning rate η.

3.2 FBU Depression Relevance and Video-level Loss Computation (FVC)

As illustrated in Fig. 3, our FVC phase is achieved by three main steps: (1) **Varying-length Facial Behaviour Segment Proposal (FBSP)** which identifies varying-length facial behaviour segments (i.e., they are made up of different numbers of consecutive FBUs); (2) **Depression-relevance Score Computation (DSC)** which evaluates the contribution of every FBU to the video-level depression status according to its associated facial behaviour segments and the segments' contributions in reflecting the target video-level depression status; and (3) video-level loss computation (VLC) which computes the video-level loss by considering the top-ranked depression-informative facial behaviour segments of the given face video.

Varying-length facial behaviour segment proposals (FBSP): Since depression-related facial behaviours vary widely in their temporal durations (Xu et al., 2024), this step constructs depression-informative facial behaviour segments of variable temporal lengths, denoted as $\{S_t\}_{t=1}^{T}$. Let

$\tau(t) = \max(1, \lfloor t - d_t/2 \rceil), \gamma(t) = \min(T, \lfloor t + d_t/2 \rceil)$, each $S_t = \{u_{\tau(t)} : u_{\gamma(t)}\}$ is centred at FBU u_t with its temporal duration d_t predicted by an MLP, where rounding $\lfloor \cdot \rceil$ is used to ensure integer FBU indexing. To compute d_t, following anchor-based temporal action localisation strategies (Liu & Liu, 2025; Reza et al., 2024), we pre-define a set of segment-length candidates $\mathcal{B} = \{b_m\}_{m=1}^{M}$, using $\mathcal{B} = \{2, 3, 4, 6, 8, 12\}$ throughout so that segments span roughly 2 to 12s. (Longer candidates degrade performance, as detailed in Appendix Sec. E). Given the corresponding FBU feature $\mathbf{f}_t$ derived from u_t, the MLP outputs a probability distribution over $\mathcal{B}$, resulting in the predicted segment length d_t as:

$$d_t = \sum_{m=1}^{M} [\mathrm{softmax}(\mathrm{MLP}_{\text{len}}(\mathbf{f}_t))]_m b_m. \tag{3}$$

Then, all FBU features $F_t = \{\mathbf{f}_{\tau(t)} : \mathbf{f}_{\gamma(t)}\}$ within each S_t are aggregated as a unified facial behaviour segment-level representation $\mathbf{s}_t \in \mathbb{R}^D$ as:

$$\mathbf{s}_t = \mathrm{Agg}(F_t) \tag{4}$$

where $\mathrm{Agg}(\cdot)$ computes the element-wise mean $\mathbf{s}_t^{\text{mean}}$ and maximum $\mathbf{s}_t^{\text{max}}$ across FBU features within S_t, together with an attentive vector $\mathbf{s}_t^{\text{att}}$, a weighted sum of the FBU features $\{\mathbf{f}_i\}_{i=\tau(t)}^{\gamma(t)}$ within S_t, with weights obtained by applying a softmax over MLP-predicted scores for the FBUs in S_t. We concatenate the mean vector $\mathbf{s}_t^{\text{mean}}$, maximum vector $\mathbf{s}_t^{\text{max}}$ and attentive vector $\mathbf{s}_t^{\text{att}}$, then project the concatenated vector through another MLP to obtain the D-dimensional segment representation (SR) $\mathbf{s}_t$. To train the duration predictor despite boundary rounding, we approximate its gradients using differentiable soft-window aggregation (Appendix Sec. G). Appendix Sec. D compares adaptive FBSP durations with fixed and shared learned widths.

Depression-relevance Score Computation (DSC): Based on the SRs learned via FBSP, the DSC assigns a depression relevance score to every FBU. Specifically, it first ranks facial behaviour segments using a depression category-guided strategy (Liao et al., 2026), achieved by two MLP branches. As illustrated in Fig. 3 (DSC), both MLPs are learned to separately predict a depression category vector (i.e., a 4-element vector indicating predictions on none, mild, modest and severe depression categories) from each facial segment representation (SR) $\mathbf{s}_t$, resulting in two category matrices $\mathcal{H}_1, \mathcal{H}_2 \in \mathbb{R}^{T\times 4}$. Particularly, the first MLP branch applies softmax row-wisely over $\mathcal{H}_1$ to produce a probability distribution over the four depression categories for each SR, resulting in a matrix $\hat{\mathcal{H}}_1$, while the second MLP applies a softmax column-wisely over the $\mathcal{H}_2$ to produce normalised relevance scores for all SRs with respect to each depression category, resulting in a matrix $\hat{\mathcal{H}}_2$ as:

$$\hat{\mathcal{H}}_1 = \underset{\text{row}}{\mathrm{softmax}}\left(\mathrm{MLP}_1(\{\mathbf{s}_t\}_{t=1}^{T})\right), \qquad \hat{\mathcal{H}}_2 = \underset{\text{column}}{\mathrm{softmax}}\left(\mathrm{MLP}_2(\{\mathbf{s}_t\}_{t=1}^{T})\right). \tag{5}$$

Here, the element of t-th row and c-th column in $\hat{\mathcal{H}}_1$ and $\hat{\mathcal{H}}_2$ represent the probability of the t-th facial behaviour segment indicating the c-th depression category, and the normalised relevance between the t-th facial behaviour segment and the c-th depression category in the context of all SRs, respectively. Then, we conduct element-wise multiplication between $\hat{\mathcal{H}}_1$ and $\hat{\mathcal{H}}_2$, outputting a matrix $\mathcal{P}$ whose element is a weighted relevance score describing the association between the corresponding facial behaviour segment and a specific depression category. Here, we compute column-wise sum of $\mathcal{P}$, where the depression category corresponding to the c-th column with the maximum sum value is assumed as the final video-level depression category prediction $\hat{c}$. It then ranks all facial behaviour segments by their relevance scores in the $\hat{c}$-th column of $\mathcal{P}$, and select the top-K segments (e.g., top 40%) as $S^{\text{sel}} = \{\mathbf{s}_{\pi_1}, \cdots, \mathbf{s}_{\pi_K}\}$:

$$\hat{c} = \arg\max_{c} \sum_{t=1}^{T} \mathcal{P}_{t,c}, \qquad \mathbf{r} = \mathcal{P}_{:,\hat{c}}, \qquad (\pi_1, \ldots, \pi_K) = \mathrm{TopK}(\mathbf{r}, K). \tag{6}$$

Since FBSP centres a segment at every FBU, the selected segments can overlap. We compute each FBU's relevance by summing the scores of all selected segments containing it,

$$[\mathbf{u}]_t = \sum_{k=1}^{K} \mathbf{1}_{\{\,\tau(\pi_k)\le t\le \gamma(\pi_k)\,\}}\, [\mathbf{r}]_{\pi_k}, \quad t = 1, \ldots, T, \tag{7}$$

where π_k indexes the k-th selected segment, $\tau(\pi_k)$ and $\gamma(\pi_k)$ are its start and end FBU indices, and the indicator equals 1 when that segment contains FBU u_t and 0 otherwise. An FBU covered by multiple selected segments therefore accumulates their scores, while an uncovered FBU receives zero relevance. Each FBU accumulates a subset of the non-negative relevance scores, whose total is at most one, ensuring $[\mathbf{u}]_t \in [0, 1]$.

Video-level loss computation (VLC): Based on the selected segment representations S^{sel}, a video-level descriptor $\mathbf{z}$ is achieved by conducting relevance score-weighted sum of them, which is then fed to an MLP for jointly predicting both the video-level depression severity score $\hat{y}$ and a $(C-1)$-dimensional vector $\hat{\mathbf{q}}$ of cumulative ordinal logits as:

$$\mathbf{z} = \sum_{k=1}^{K} [\mathbf{r}]_{\pi_k} \mathbf{s}_{\pi_k} \qquad (\hat{y}, \hat{\mathbf{q}}) = \text{MLP}_{\text{pre}}(\mathbf{z}). \tag{8}$$

Given the depression label y, its depression category label c and a cumulative binary target vector $\mathbf{q}$ derived from c, the full video-level loss is computed as:

$$\mathcal{L}_{\text{vid}} = \mathcal{L}_{\text{reg}}(\hat{y}, y) + \mathcal{L}_{\text{ord}}(\hat{\mathbf{q}}, \mathbf{q}) + \mathcal{L}_{\text{cls}}(\mathbf{a}, c), \tag{9}$$

where $\mathcal{L}_{\text{reg}}(\cdot)$ is mean squared error (MSE) loss measuring the predicted depression severity; $\mathcal{L}_{\text{ord}}(\cdot)$ is an ordinal classification loss; and $\mathcal{L}_{\text{cls}}(\cdot)$ is the cross entropy loss, where both $\mathcal{L}_{\text{ord}}(\cdot)$ and $\mathcal{L}_{\text{cls}}(\cdot)$ measure the predicted depression category. We sum the three loss terms with a weight of 1 for each term. Here, $\mathbf{a}$ denotes the MIL category scores, and Appendix Sec. H details both classification losses and their target encoding. The video-level loss $\mathcal{L}_{\text{vid}}$ together with the per-FBU relevance vector $\mathbf{u}$, constitutes the outputs of FVC in Eq. 1 that directly drives the subsequent gradient modulation.

3.3 Gradient Modulation and Video-level Loss Back-propagation

The **GMVB** phase uses the depression relevance estimated from variable-length facial behaviour segments to modulate backbone learning. FVC weights segment representations for video-level prediction, while GMVB uses the per-FBU relevance to weight their contributions to the backbone update. We first back-propagate the video-level loss through FVC to compute its gradient with respect to each FBU feature $\mathbf{f}_t$ (Step 2.1 in Fig. 2):

$$\mathbf{g}_t^{\mathbf{f}} = \partial \mathcal{L}_{\text{vid}} / \partial \mathbf{f}_t \tag{10}$$

These FVC gradients provide video-level supervision for the FBU feature extractor $\mathcal{M}$. We scale each gradient by $[\mathbf{u}]_t$, which aggregates the relevance scores of selected segments containing u_t (Eq. 7). Back-propagation through the backbone requires the intermediate features and computation graph discarded during FVC feature extraction. We therefore pass each FBU through $\mathcal{M}$ again, with θ unchanged, to reconstruct backbone intermediate feature $\mathbf{h}_t = \mathcal{M}_{\text{int}}(u_t; \theta)$ and the backbone computation graph for that FBU. We then apply the chain rule to compute each FBU's relevance-weighted contribution to the parameter gradients of $\mathcal{M}$ (Step 2.2 in Fig. 2):

$$\tilde{\mathbf{g}}_t^{(\theta)} = \frac{\partial \mathbf{h}_t}{\partial \theta}^{\top} \frac{\partial \mathbf{f}_t}{\partial \mathbf{h}_t}^{\top} \left([\mathbf{u}]_t \mathbf{g}_t^{\mathbf{f}} \right) = [\mathbf{u}]_t \, \mathbf{g}_t^{(\theta)}, \tag{11}$$

where $\mathbf{g}_t^{(\theta)}$ is the unmodulated backbone gradient induced by u_t. We scale each FBU's feature gradient by its depression relevance score $[\mathbf{u}]_t$ computed by DSC. A smaller score attenuates that FBU's contribution, while a zero score removes it from the backbone update. Finally, we update the model parameters θ using the sum of the relevance-weighted gradient contributions from all FBUs in the video:

$$\theta \leftarrow \text{Update}(\theta, \sum_{t=1}^{T} \tilde{\mathbf{g}}_t^{(\theta)}, \eta) \tag{12}$$

where $\text{Update}(\cdot)$ denotes the model weight update performed by the chosen optimiser (e.g., AdamW (Loshchilov & Hutter, 2017)) with learning rate η.

This update combines full-video supervision with explicit depression relevance weighting at the feature-learning stage. The implementation integrates model-activation recomputation as in gradient checkpointing (Chen et al., 2016) and feature-gradient caching as in GradCache (Gao et al.,

2021). We cache the feature gradients and release each FBU's computed intermediate features and computation graph after accumulating its relevance-weighted parameter gradients. This limits backbone activation storage to one FBU, while feature and FVC storage still grows with video length. Appendix Sec. I distinguishes the relevance-weighted update from ordinary loss gradients and explains its sequential computation. Training efficiency is analysed in Appendix Sec. J.

4 EXPERIMENTS

We evaluate our approach on the AVEC 2013 and 2014 (Valstar et al., 2013; 2014) datasets. Participants in both AVEC 2013 and AVEC 2014 datasets are labelled with BDI-II depression scores ranging from 0 to 63. The AVEC 2013 recorded a total of 150 video clips from 82 participants, with their lengths between 20 and 50 mins, while the AVEC 2014 contains two subsets: Northwind and Freeform, where each contains 150 clips ranging from 6 seconds to 4 minutes. Following previous studies (Niu et al., 2024b; He et al., 2025; Xu et al., 2024) root mean square error (RMSE) and mean absolute error (MAE) are measured. More data preprocessing, implementation and metric details are provided in Appendix Sec. A.

4.1 COMPARISON WITH EXISTING METHODS AND ABLATION STUDY

Table 1 compares the performance of Video Action Transformer (VAT) (Girdhar et al., 2019) and LMTformer (LMT) (He et al., 2025) trained under our proposed DMVS strategy with prior representative ADA approaches on AVEC datasets. For our AVEC 2014 results in Table 1 and 3, we average the Freeform and Northwind predictions for the same subject before computing RMSE and MAE, whereas Table 2 evaluates each subset separately. On AVEC 2013, the VAT achieves the best overall performance, reaching an RMSE of 6.88 and an MAE of 5.17, which improves upon the strongest prior video-level model (Wang et al. (2025b), 6.99/5.71). On AVEC 2014, the VAT attains the lowest MAE of 4.97, improving upon the MAE of 5.11 reported by Liao et al. (2026). Our video-level DMVS strategy also improves the lightweight LMT compared with its original segment-level training (He et al., 2025). It reduces RMSE/MAE from 7.75/6.12 to 7.25/5.91 on AVEC 2013 and from 7.97/6.05 to 6.94/5.59 on AVEC 2014, outperforming a range of representative ADA methods.

We conduct ablation studies on AVEC 2013 and AVEC 2014 (Freeform and Northwind) using VAT and LMT to evaluate FVC and GMVB modules. Within each dataset block in Table 2, the first row is a segment-level (Seg.) baseline. Each segment is trained to predict the video depression label independently, and the segment predictions are averaged to obtain the video prediction. Rows 2–6 report results from end-to-end training with video-level (Vid.) supervision. In Row 2, we obtain an individual prediction from each FBU feature and average these predictions to form the video-level prediction, using GradCache for video-level training without FBSP, DSC or GMVB. Rows 2–5 set $[\mathbf{u}]_t = 1$, which is equivalent to using GradCache (Gao et al., 2021) for backbone updates. Row 5 uses the full FVC. Row 6 adds GMVB module, yielding the full DMVS. Video-level training with GradCache alone does not consistently outperform segment-level training. Across datasets, either FBSP or DSC under video-level supervision improves over the segment-level baseline, and their combination yields larger gains. For VAT, adding both FVC components reduces RMSE from 7.47 to 7.14 on AVEC 2013, from 7.79 to 7.52 on Freeform, and from 7.69 to 7.31 on Northwind, while adding GMVB further reduces the errors to 6.88, 7.19, and 7.09, respectively. LMT follows the same trend. Since Rows 5 and 6 share the same FVC, video-level loss terms and gradient-caching procedure, the further RMSE reductions isolate the benefit of GMVB relevance weighting over unmodulated GradCache updates. Together, these results support the complementary benefits of FVC segment modelling and GMVB relevance weighting under video-level supervision. Appendix Sec. C compares GMVB with unit-weight, mean-weight and shuffled-weight gradient settings.

4.2 GENERALISATION VALIDATION

We apply DMVS to five representative spatio-temporal models, LMT (He et al., 2025), MDN (de Melo et al., 2021), TSSTANet (Pan et al., 2024), X3D (Feichtenhofer, 2020), and VAT (Girdhar et al., 2019), on the AVEC 2013 and 2014 datasets to verify its effectiveness and generality. LMT, MDN, and TSSTANet are ADA-specific segment-level models, whereas X3D and VAT are general-purpose temporal CNN and Transformer models. For a clear comparison, we report three settings for

Table 1: ADA results on AVEC 2013/2014. Lower RMSE and MAE are better.

Methods	AVEC 2013		AVEC 2014	
	RMSE	MAE	RMSE	MAE
Al Jazaery & Guo (2018)	9.28	7.37	9.20	7.22
Uddin et al. (2020)	8.93	7.04	8.78	6.86
De Melo et al. (2020a)	7.90	5.98	7.61	5.82
He et al. (2021)	8.39	6.59	8.30	6.51
Shang et al. (2023)	8.20	6.38	7.84	6.08
Song et al. (2022)	8.10	6.16	7.15	5.95
Uddin et al. (2022)	7.32	5.90	6.98	5.75
Pan et al. (2023)	7.26	5.97	7.30	5.99
Niu et al. (2024b)	7.49	5.43	7.27	5.63
Xu et al. (2024)	7.57	5.95	7.18	5.86
Li et al. (2025)	8.64	6.82	8.24	6.65
Kumar et al. (2025)	7.89	6.15	8.11	6.14
He et al. (2025)	7.75	6.12	7.97	6.05
Wu et al. (2025)	7.08	5.41	6.62	5.45
Wang et al. (2025b)	6.99	5.71	6.93	5.36
Liao et al. (2026)	7.48	5.77	6.89	5.11
Ours LMT	7.25	5.91	6.94	5.59
Ours VAT	**6.88**	**5.17**	**6.61**	**4.97**

Table 2: FVC/GMVB ablation with VAT and LMT. Bold marks the lowest RMSE.

Dataset	Sup.	FVC		GMVB	RMSE↓	
		FBSP	DSC		VAT	LMT
AVEC 2013	Seg.				7.47	7.74
	Vid.				7.60	7.66
	Vid.	✓			7.40	7.71
	Vid.		✓		7.41	7.56
	Vid.	✓	✓		7.14	7.39
	Vid.	✓	✓	✓	**6.88**	**7.25**
Freeform	Seg.				7.79	7.69
	Vid.				7.83	7.70
	Vid.	✓			7.77	7.63
	Vid.		✓		7.60	7.62
	Vid.	✓	✓		7.52	7.47
	Vid.	✓	✓	✓	**7.19**	**7.40**
Northwind	Seg.				7.69	7.84
	Vid.				7.62	7.89
	Vid.	✓			7.60	7.76
	Vid.		✓		7.52	7.76
	Vid.	✓	✓		7.31	7.68
	Vid.	✓	✓	✓	**7.09**	**7.54**

Table 3: Applying our video-level training strategy (DMVS) to five existing models used for ADA. Reproduced and DMVS augmented RMSE results on AVEC 2013 and 2014 datasets are compared.

Methods	Original		Reproduced		DMVS	
	2013	2014	2013	2014	2013	2014
LMT He et al. (2025)	7.75	7.97	7.74	7.30	**7.25**	**6.94**
MDN de Melo et al. (2021)	8.13	8.16	8.73	8.32	8.54	**8.00**
TSSTANet Pan et al. (2024)	9.65	9.55	9.33	8.97	**9.04**	**8.61**
X3D Feichtenhofer (2020)	–	–	9.36	8.93	**9.18**	**8.70**
VAT Girdhar et al. (2019)	–	–	7.47	7.22	**6.88**	**6.61**

each model: **Original** results from the papers, our **Reproduced** results obtained by re-training/re-implementing the baselines in our codebase, and **DMVS** results applying our video-level training strategy. We use the same DMVS hyperparameter settings across all five backbones. As shown in Table 3, our reproduced LMT and TSSTANet achieve lower RMSE than originally reported, while our reproduced MDN has higher RMSE. We use the reproduced results as the baselines for evaluating the effect of DMVS within the same code base. Comparing DMVS results to reproduced ones, the DMVS yields consistent gains, suggesting that DMVS is a model-agnostic and plug-and-play training strategy that can improve face video-based depression assessment across diverse backbones. Results for the individual AVEC 2014 subsets are reported in Appendix Table 4. We further evaluate DMVS on personality analysis using ChaLearn 2016, with the experimental setup and results reported in Appendix Sec. L.

5 Conclusion

We introduce DMVS for end-to-end full-video depression assessment, combining variable-length facial behaviour segment modelling with depression relevance-weighted gradient modulation. This enables depression relevance estimated from variable-length facial behaviour segments to modulate backbone gradients under full-video supervision. Experiments validate DMVS as a model-agnostic and plug-and-play training strategy.

Ethics Statement

We study video-based prediction of self-reported depression severity on existing benchmarks. The results do not establish clinical validity or equitable performance across demographic groups. Face videos and inferred depression severity scores are sensitive, and any use beyond benchmark research should require explicit informed consent, privacy safeguards and independent validation. The method should not be used for diagnosis or consequential screening. Appendix Sec. M provides the full discussion of limitations and ethical considerations.

Reproducibility Statement

We describe the FVC and GMVB procedures in Secs. 3.2 and 3.3. Appendix Sec. A specifies the datasets, preprocessing, training settings, evaluation metrics and checkpoint-selection protocol. Appendix Secs. G and H detail the duration-gradient approximation and supervision targets, while Appendix Sec. I states the assumptions for gradient recomputation and distinguishes the modulated update from the ordinary loss gradient. Appendix Sec. E reports the shared hyperparameter settings and sensitivity analyses, and Appendix Sec. J describes the hardware and memory-profiling procedure.

AI Use Statement

We used generative AI tools to obtain feedback on the appropriateness of our experimental design and to polish the language and improve the clarity of the manuscript. We also used generative AI tools to assist with mathematical derivations concerning hard segment boundaries and gradient backpropagation. We did not use generative AI tools to formulate or refine hypotheses or to generate, clean or reformat datasets. We checked the AI-assisted derivations for mathematical correctness, assessed the experimental design feedback, and reviewed the language revisions for technical accuracy and consistency with the methods and results. We take responsibility for the final content of this paper.

References

Mohamad Al Jazaery and Guodong Guo. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing*, 12(1):262–268, 2018.

Lucinda Archer, Kym I. E. Snell, Joie Ensor, Mohammed T. Hudda, Gary S. Collins, and Richard D. Riley. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*, 40(1):133–146, 2021. doi: 10.1002/sim.8766.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018. doi: 10.1109/FG.2018.00019.

Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2846–2854, 2016.

Hugo Bohy, Minh Tran, Kevin El Haddad, Thierry Dutoit, and Mohammad Soleymani. Social-mae: A transformer-based multimodal autoencoder for face and voice. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5. IEEE, 2024.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Fernando De la Torre, Joan Campoy, Zara Ambadar, and Jeffrey F Cohn. Temporal segmentation of facial behavior. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Wheidima Carneiro De Melo, Eric Granger, and Abdenour Hadid. Depression detection based on deep distribution learning. In *2019 IEEE international conference on image processing (ICIP)*, pp. 4544–4548. IEEE, 2019.

Wheidima Carneiro De Melo, Eric Granger, and Abdenour Hadid. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE transactions on affective computing*, 13(3):1581–1592, 2020a.

Wheidima Carneiro De Melo, Eric Granger, and Miguel Bordallo Lopez. Encoding temporal information for automatic depression recognition from facial analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1080–1084. IEEE, 2020b.

Wheidima Carneiro de Melo, Eric Granger, and Miguel Bordallo Lopez. Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE transactions on affective computing*, 14(1):578–590, 2021.

Heiner Ellgring. *Non-verbal communication in depression*. Cambridge University Press, 2007.

Huiting Fan, Xingnan Zhang, Yingying Xu, Jiangxiong Fang, Shiqing Zhang, Xiaoming Zhao, and Jun Yu. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Information Fusion*, 104:102161, 2024.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213, 2020.

Yihang Feng, Zequn Guan, Chao Che, and Lizhong Liang. CSTEformer: A cross-scale spatiotemporal perception enhanced transformer for depression detection in asian populations. *Biomedical Signal Processing and Control*, 117:109604, 2026. doi: 10.1016/j.bspc.2026.109604.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pp. 316–321, 2021.

Jean-Guido Gehricke and David Shapiro. Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Research*, 95(2):157–167, 2000.

Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.

Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 244–253, 2019.

Meng-Meng Han, Xing-Yun Li, Xin-Yu Yi, et al. Automatic recognition of depression based on audio and video: A review. *World Journal of Psychiatry*, 14(2):225–233, 2024.

Lang He, Jonathan Cheung-Wai Chan, and Zhongmin Wang. Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing*, 422:165–175, 2021.

Lang He, Zheng Li, Prayag Tiwari, Cui Cao, Jize Xue, Feng Zhu, and Di Wu. Depressformer: Leveraging video swin transformer and fine-grained local features for depression scale estimation. *Biomedical Signal Processing and Control*, 96:106490, 2024.

Lang He, Junnan Zhao, Jie Zhang, Jiewei Jiang, Senqing Qi, Zhongmin Wang, and Di Wu. Lmtformer: facial depression recognition with lightweight multi-scale transformer from videos. *Applied Intelligence*, 55(3):195, 2025.

Lang He, Kai Chen, Junnan Zhao, Yimeng Wang, Ercheng Pei, Haifeng Chen, Jiewei Jiang, Shiqing Zhang, Jie Zhang, Zhongmin Wang, Tao He, and Prayag Tiwari. LMVD: A large-scale multimodal vlog dataset for depression detection in the wild. *Information Fusion*, 126:103632, 2026. doi: 10.1016/j.inffus.2025.103632.

Jayson Jeganathan, Megan Campbell, Matthew Hyett, Gordon Parker, and Michael Breakspear. Quantifying dynamic facial expressions under naturalistic conditions. *Elife*, 11:e79581, 2022. doi: 10.7554/eLife.79581.

Fernando C Krause, Eftihia Linardatos, David M Fresco, and Michael T Moore. Facial emotion recognition in major depressive disorder: A meta-analytic review. *Journal of affective disorders*, 293:320–328, 2021.

Puneet Kumar, Shreshtha Misra, Zhuhong Shao, Bin Zhu, Balasubramanian Raman, and Xiaobai Li. Multimodal interpretable depression analysis using visual, physiological, audio and textual data. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5305–5315. IEEE, 2025.

Yonghong Li, Shan Qu, and Xiuzhuang Zhou. Conformal depression prediction. *IEEE Transactions on Affective Computing*, 2025.

Yunan Li, Jun Wan, Qiguang Miao, Sergio Escalera, Huijuan Fang, Huizhou Chen, Xiangda Qi, and Guodong Guo. Cr-net: A deep classification-regression network for multimodal apparent personality analysis. *International Journal of Computer Vision*, 128(12):2763–2780, 2020.

Yutong Li, Zhenyu Liu, Li Zhou, Xiaoyan Yuan, Zixuan Shangguan, Xiping Hu, and Bin Hu. A facial depression recognition method based on hybrid multi-head cross attention network. *Frontiers in Neuroscience*, 17:1188434, 2023. doi: 10.3389/fnins.2023.1188434.

Rongfan Liao, Siyang Song, and Hatice Gunes. An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition. *IEEE Transactions on Affective Computing*, 15(3):1590–1607, 2024.

Rongfan Liao, Xiangyu Kong, Shiqing Tang, Lang He, Changzeng Fu, Weicheng Xie, Xiaofeng Liu, Lu Liu, and Siyang Song. Explainable depression assessment from face videos by weakly supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1928–1936, 2026. doi: 10.1609/aaai.v40i3.37173.

Zhiqiang Liu, Gengjia Zhang, Jisun Hong, Jihun Lee, Daegil Choi, Jaehyo Jung, and Meina Li. DepSTDSAM: Spatio-temporal dual-sparse attention model for depression prediction based on facial behavioral structure representations. *Neurocomputing*, 682:133449, 2026. doi: 10.1016/j.neucom.2026.133449.

Ziyi Liu and Yangcen Liu. Bridge the gap: From weak to full supervision for temporal action localization with pseudoformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8711–8720, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Kaining Mao, Yuqi Wu, and Jie Chen. A systematic review on automated clinical depression diagnosis. *npj Mental Health Research*, 2:20, 2023.

Mingyue Niu, Jianhua Tao, Bin Liu, Jian Huang, and Zheng Lian. Multimodal spatiotemporal representation for automatic depression level detection. *IEEE transactions on affective computing*, 14(1):294–307, 2020.

Mingyue Niu, Jianhua Tao, and Bin Liu. Multi-scale and multi-region facial discriminative representation for automatic depression level prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1325–1329. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9413504.

Mingyue Niu, Ming Li, and Changzeng Fu. Pointtransform networks for automatic depression level prediction via facial keypoints. *Knowledge-Based Systems*, 297:111951, 2024a.

Mingyue Niu, Ya Li, Jianhua Tao, Xiuzhuang Zhou, and Björn W Schuller. Depressionmlp: A multi-layer perceptron architecture for automatic depression level prediction via facial keypoints and action units. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024b.

Yuchen Pan, Yuanyuan Shang, Zhuhong Shao, Tie Liu, Guodong Guo, and Hui Ding. Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. *IEEE Transactions on Affective Computing*, 2023.

Yuchen Pan, Yuanyuan Shang, Tie Liu, Zhuhong Shao, Guodong Guo, Hui Ding, and Qiang Hu. Spatial–temporal attention network for depression recognition from facial videos. *Expert systems with applications*, 237:121410, 2024.

Smitha Joyce Pinto and Mimmo Parente. Comprehensive review of depression detection techniques based on machine learning approach. *Soft Computing*, 28(17):10701–10725, 2024.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 400–418. Springer, 2016.

Sakib Reza, Yuexi Zhang, Mohsen Moghaddam, and Octavia Camps. Hat: History-augmented anchor transformer for online temporal action localization. In *European Conference on Computer Vision*, pp. 205–222. Springer, 2024.

Jonathan Rottenberg. Emotions in depression: What do we really know? *Annual Review of Clinical Psychology*, 13:241–263, 2017.

Yuanyuan Shang, Yuchen Pan, Xiao Jiang, Zhuhong Shao, Guodong Guo, Tie Liu, and Hui Ding. Lqgdnet: A local quaternion and global deep network for facial depression recognition. *IEEE Transactions on Affective Computing*, 14(3):2557–2563, 2023. doi: 10.1109/TAFFC.2021.3139651.

Haotian Shen, Siyang Song, and Hatice Gunes. Multi-modal human behaviour graph representation learning for automatic depression assessment. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–10. IEEE, 2024. doi: 10.1109/FG59268.2024.10581964.

Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 158–165. IEEE, 2018.

Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2022.

Eriko Sugimori and Mayu Yamaguchi. Subthreshold depression is associated with altered facial expression and impression formation via subjective ratings and action unit analysis. *Scientific reports*, 15(1):30761, 2025.

Md Azher Uddin, Joolekha Bibi Joolee, and Young-Koo Lee. Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. *IEEE Transactions on Affective Computing*, 13(2):864–870, 2020.

Md Azher Uddin, Joolekha Bibi Joolee, and Kyung-Ah Sohn. Deep multi-modal network based automated depression severity estimation. *IEEE transactions on affective computing*, 2022.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 3–10, 2013.

Michel Valstar, Björn Schuller, Roddy Cowie, Maja Pantic, and Fabien Ringeval. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10. ACM, 2014. doi: 10.1145/2661806.2661807.

Jialou Wang, Honglei Li, Wai Lok Woo, and Shan Shan. A single modality apparent first impression personality recognition model with temporal emotion based lstm. *Expert Systems with Applications*, 259:125114, 2025a.

Yaowei Wang, Zulong Lin, Chengrong Yang, Yujue Zhou, and Yun Yang. Automatic depression recognition with an ensemble of multimodal spatio-temporal routing features. *IEEE Transactions on Affective Computing*, 2025b.

Zijian Wu, Leijing Zhou, Shuanglin Li, Changzeng Fu, Jun Lu, Jing Han, Yi Zhang, Zhuang Zhao, and Siyang Song. Depmgnn: Matrixial graph neural network for video-based automatic depression assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1610–1619, 2025.

Jiaqi Xu, Hatice Gunes, Keerthy Kusumam, Michel Valstar, and Siyang Song. Two-stage temporal modelling framework for video-based depression recognition using graph representation. *IEEE Transactions on Affective Computing*, 2024.

Keshan Yan, Shengfa Miao, Xin Jin, Yongkang Mu, Hongfeng Zheng, Yuling Tian, Puming Wang, Qian Yu, and Da Hu. Tcedn: A lightweight time-context enhanced depression detection network. *Life*, 14(10):1313, 2024.

Shiqing Zhang, Xingnan Zhang, Xiaoming Zhao, Jiangxiong Fang, Mingyue Niu, Ziping Zhao, Jun Yu, and Qi Tian. Mtdan: A lightweight multi-scale temporal difference attention networks for automated video depression detection. *IEEE transactions on affective computing*, 15(3):1078–1089, 2023.

Xiaoming Zhao, Yuehui Liao, Zhiwei Tang, Yicheng Xu, Xin Tao, Dandan Wang, Guoyu Wang, and Hongsheng Lu. Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning. *Frontiers in Neuroscience*, 16:1107284, 2023.

Jinzhao Zhou, Xingming Zhang, Yang Liu, and Xiangyuan Lan. Facial expression recognition using spatial-temporal semantic graph network. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1961–1965. IEEE, 2020.

Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE transactions on affective computing*, 11(3):542–552, 2018.

A Datasets and implementation details

Dataset: AVEC 2013 (Valstar et al., 2013) and AVEC 2014 (Valstar et al., 2014) remain the only publicly available benchmark datasets for face video-based automatic depression assessment **with standardised depression severity annotations**. The limited availability of such datasets is a common constraint in this research area, and almost all prior face video-based methods for depression severity detection have therefore been evaluated on these benchmarks. To further examine the generalisation of DMVS, we also evaluate it on the related ChaLearn 2016 First Impressions personality dataset (Ponce-López et al., 2016).

The AVEC 2013 recorded a total of 150 video clips from 82 participants, with their lengths between 20 and 50 mins, while the AVEC 2014 contains two subsets: Northwind and Freeform, where each contains 150 clips ranging from 6 seconds to 4 minutes. Both AVEC 2013 and 2014 datasets separate 50 videos for training, 50 videos for validation and 50 videos for testing. The ChaLearn 2016 dataset contains 10,000 short face videos collected from YouTube, each lasting about 15 seconds, and provides annotations for the Big-Five apparent personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Each trait is annotated as a continuous score in $[0, 1]$ according to human judgements. Following prior work (Li et al., 2020), we discretise each trait label into four bins for category-guided segment ranking and train one model for each personality trait.

Evaluation protocol: For AVEC 2013 and AVEC 2014, we follow the training and checkpoint-selection protocol implemented in prior depression assessment studies (He et al., 2025; Wu et al., 2025; Liao et al., 2026): we merge the official training and development splits for training, select the checkpoint with the lowest RMSE on the official test split, and report its test metrics. The combined training split is also used in AVEC 2014 experiments by Li et al. (2023). Because checkpoint selection uses test labels, these metrics are subject to selection bias and should not be interpreted as estimates from a held-out test set. We use this protocol for comparison with published test-set results.

Implementation details: In our experiments, we split each video into consecutive, non-overlapping 32-frame segments, each termed a facial behaviour unit (FBU). We extract the face image from each frame using OpenFace 2 (Baltrusaitis et al., 2018) and resize it to 224×224 pixels. We apply random flipping, colour jittering and Gaussian blurring for data augmentation. We train for a maximum of 50 epochs. During the first five epochs, we increase the learning rate from 1×10^{-8} to 1×10^{-4} for warm-up. Over the remaining 45 epochs, we use cosine annealing to reduce the learning rate to 1×10^{-6}. AdamW is uniformly used as the optimiser in all experiments. For the ablation study in Table 2, we train the segment-level baseline with the MSE regression loss $\mathcal{L}_{\mathrm{reg}}$ and ordinal classification loss $\mathcal{L}_{\mathrm{ord}}$. When DSC is included, we additionally use the categorical cross-entropy term $\mathcal{L}_{\mathrm{cls}}$ required by its category-guided segment ranking.

Metrics: We follow previous studies (Valstar et al., 2013; 2014; Song et al., 2022; Niu et al., 2024b; de Melo et al., 2021; Xu et al., 2024) to measure root mean square error (RMSE), mean absolute error (MAE), Pearson Correlation Coefficient (PCC), and Concordance Correlation Coefficient (CCC) on the AVEC depression datasets. Given N samples with ground-truth depression scores y_i and predictions $\hat{y}_i$, we define:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \qquad \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|, \tag{13}$$

$$\mathrm{PCC} = \rho = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}}, \tag{14}$$

$$\mathrm{CCC} = \rho_c = \frac{2\rho\,\sigma_{\hat{y}}\,\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + \left(\bar{\hat{y}} - \bar{y}\right)^2}. \tag{15}$$

Here, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$ are sample means; σ_y and $\sigma_{\hat{y}}$ are sample standard deviations with $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$ and $\sigma_{\hat{y}}^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2$.

For ChaLearn 2016, we report the standard personality prediction accuracy for each trait and the mean accuracy over the five traits. For trait m, with ground-truth personality scores $s_i^{(m)} \in [0, 1]$ and predictions $\hat{s}_i^{(m)}$, the accuracy is computed as

$$\mathrm{Acc}^{(m)} = 1 - \frac{1}{N}\sum_{i=1}^{N}\left|\hat{s}_i^{(m)} - s_i^{(m)}\right|, \tag{16}$$

and reported as a percentage. The mean accuracy is computed as $\frac{1}{5}\sum_{m=1}^{5}\mathrm{Acc}^{(m)}$.

B Results on the AVEC 2014 subsets

Table 4 compares the reproduced models with their DMVS counterparts on Freeform and Northwind.

C Effect of GMVB gradient modulation

We evaluate the effect of FBU gradient modulation using VAT on AVEC 2014 Northwind. All variants retain the same FVC, video-level losses and two-pass gradient-caching procedure. They differ

Table 4: Applying DMVS to five existing models used for ADA on the AVEC 2014 Freeform and Northwind subsets. We compare reproduced and DMVS results using RMSE and MAE (lower is better). Bold indicates the better result for each model and metric within each subset.

Methods	Freeform				Northwind			
	Reproduced		DMVS		Reproduced		DMVS	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
LMT He et al. (2025)	7.69	6.06	**7.40**	**5.79**	7.84	6.37	**7.54**	**6.15**
MDN de Melo et al. (2021)	9.44	7.69	**9.25**	**7.54**	8.33	**6.51**	**8.00**	6.53
TSSTANet Pan et al. (2024)	9.46	7.79	**9.01**	**7.03**	9.43	7.29	**9.19**	**7.19**
X3D Feichtenhofer (2020)	9.50	7.24	**9.18**	**6.97**	9.31	6.72	**9.12**	**6.68**
VAT Girdhar et al. (2019)	7.79	5.90	**7.19**	**5.55**	7.69	6.39	**7.09**	**5.35**

Table 5: Effect of GMVB gradient modulation with VAT on AVEC 2014 Northwind. Each row summarises the best RMSE from each of five experiment runs. SD denotes the sample standard deviation across runs, and π is a random permutation of FBU indices within a video. All variants retain FVC and two-pass gradient caching.

Gradient modulation	Multiplier	Minimum RMSE	Mean RMSE $\pm$ SD
Unit-weight control	1	7.3069	7.5063 ± 0.1997
Mean-weight control	$\bar{u}$	7.3412	7.5638 ± 0.2431
Shuffled-weight control	$[\mathbf{u}]_{\pi(t)}$	7.3685	7.5101 ± 0.1372
GMVB	$[\mathbf{u}]_t$	**7.0940**	$\mathbf{7.1856 \pm 0.1034}$

only in the additional multipliers applied to the cached FBU feature gradients before backbone back-propagation. Full GMVB uses the depression relevance weight $[\mathbf{u}]_t$ to modulate corresponding FBU feature gradient. The unit-weight control sets every $[\mathbf{u}]_t$ to one. The mean-weight control replaces every normalised GMVB weight by the within-video mean $\bar{u} = T^{-1} \sum_{t=1}^{T} [\mathbf{u}]_t$. The shuffled-weight control randomly permutes the weights across FBUs within each video, preserving their values while disrupting their assignment to the corresponding FBUs. None of the controls removes the relevance weighting used in the FVC forward computation. As shown in Table 5, GMVB reduces the mean RMSE by 0.3207 (4.27%) relative to unit weights and by 0.3782 (5.00%) relative to mean weights. Shuffling increases the mean RMSE from 7.1856 to 7.5101, corresponding to a 4.32% reduction for GMVB relative to the shuffled control. Its sample standard deviation is also lower in these runs. The shuffled control performs worse than GMVB, supporting the importance of assigning relevance weights to the corresponding FBUs. Together, these comparisons support the benefit of relevance-weighted FBU gradient modulation beyond the shared FVC and gradient-caching procedure.

D Effect of adaptive FBSP duration

We evaluate adaptive segment duration using VAT on AVEC 2014 Northwind. All variants retain the same segment aggregation, DSC, GMVB and training protocol. We compare fixed proposal widths of 4 and 8 FBUs, a single learned width shared across videos and addaptive FBSP for varying-length segments. The shared learned width is optimised during training and applied uniformly across all videos and temporal positions. The Adaptive FBSP row uses the same five full-DMVS runs as the GMVB row in Table 5. As shown in Table 6, adaptive FBSP reduces the mean RMSE by 0.2169 (2.93%) and 0.4709 (6.15%) relative to fixed widths of 4 and 8 FBUs, respectively, and by 0.1622 (2.21%) relative to the shared learned width. It also has the lowest sample standard deviation across these runs. These results support predicting durations from the input facial behaviours rather than using a fixed or globally shared duration.

E Parameter sensitivity analysis

We use the same fixed hyperparameter settings across all five backbones in the main AVEC experiments, including a Top-K ratio of 0.4, 32 frames per FBU and $\mathcal{B} = \{2, 3, 4, 6, 8, 12\}$. Only the

Table 6: Effect of FBSP duration with VAT on AVEC 2014 Northwind. Each row summarises the best RMSE from each of five experiment runs. SD denotes the sample standard deviation across runs. Widths are specified in FBU units before boundary rounding and clipping. All variants retain DSC and GMVB.

Duration setting	Minimum RMSE	Mean RMSE ± SD
Fixed width 4	7.2697	7.4025 ± 0.1882
Fixed width 8	7.1445	7.6565 ± 0.3810
Shared width	7.2156	7.3478 ± 0.1226
Adaptive FBSP	**7.0940**	$\mathbf{7.1856 \pm 0.1034}$

sensitivity experiments below vary these parameters, changing one parameter at a time while keeping the remaining settings fixed. These experiments assess sensitivity around the shared configuration and do not introduce backbone-specific settings into the main comparisons.

E.1 Sensitivity to the Top-K ratio

In the FVC phase, DSC ranks all candidate facial behaviour segments by their depression relevance scores, and VLC uses only the top-ranked ones for video-level prediction and loss computation. We denote the *Top-K ratio* as $K \in (0, 1]$, meaning that we select the top $\lceil KT \rceil$ segment representations (out of T candidates) according to their relevance scores. To examine sensitivity to the proportion of selected segments, we vary $K \in \{0.2, 0.4, 0.6, 0.8\}$ on three datasets while keeping all other settings fixed. As shown in Table 7, $K = 0.4$ achieves the best (or near-best) overall performance across datasets and metrics. A too-small K may discard informative depression-related segments, whereas a too-large K may introduce less relevant segments and dilute the supervision signal. Table 8 further reports the RMSE, MAE, PCC, CCC results of VAT and LMT under the default $K = 0.4$, supplementing the metrics reported in the main paper (Table 2).

Table 7: Top-K ratio sensitivity on the VAT backbone. We report RMSE/MAE/PCC/CCC on AVEC 2013 and AVEC 2014 (Freeform and Northwind) under different Top-K ratios, where K is the fraction of top-ranked segment representations in the FVC phase.

Dataset	**Portion**	**RMSE**	**MAE**	**PCC**	**CCC**
AVEC 2013	0.2	7.18	5.64	0.785	0.778
	0.4	6.88	5.17	0.804	0.800
	0.6	6.94	5.26	0.800	0.794
	0.8	7.26	5.37	0.790	0.781
Freeform	0.2	7.62	6.02	0.752	0.738
	0.4	7.19	5.55	0.783	0.773
	0.6	7.27	5.57	0.778	0.760
	0.8	7.72	5.78	0.752	0.735
Northwind	0.2	7.33	5.62	0.779	0.774
	0.4	7.09	5.35	0.794	0.788
	0.6	7.26	5.37	0.781	0.774
	0.8	7.70	5.67	0.761	0.749

Table 8: The RMSE, MAE, PCC and CCC of VAT and LMT under the chosen Top-K ratio $K = 0.4$.

Dataset	**model**	**RMSE**	**MAE**	**PCC**	**CCC**
AVEC 2013	VAT	6.88	5.17	0.804	0.800
	LMT	7.25	5.91	0.787	0.746
Freeform	VAT	7.19	5.55	0.783	0.773
	LMT	7.40	5.79	0.765	0.735
Northwind	VAT	7.09	5.35	0.794	0.788
	LMT	7.54	6.15	0.760	0.736

E.2 SENSITIVITY TO FBU LENGTH

We use 32 frames per FBU in the main experiments for all training models and examine sensitivity by varying the number of frames in each FBU among $\{16, 32, 64, 128\}$, while keeping the Top-K ratio fixed at $K = 0.4$ and leaving all other settings unchanged. As shown in Table 9, the 32-frame setting achieves the best overall performance across the three datasets. Shorter FBUs may contain insufficient temporal context for modelling depression-related facial dynamics, whereas longer FBUs may merge heterogeneous behaviours and reduce the temporal resolution of the relevance ranking.

Table 9: FBU-length sensitivity on the VAT backbone. We report RMSE/MAE/PCC/CCC on AVEC 2013 and AVEC 2014 (Freeform and Northwind) under different FBU lengths (16, 32, 64, and 128 frames), with the Top-K ratio fixed at $K = 0.4$.

Dataset	FBU(Frames)	RMSE	MAE	PCC	CCC
AVEC 2013	16	7.02	5.29	0.797	0.792
	32	6.88	5.17	0.804	0.800
	64	7.13	5.20	0.795	0.790
	128	7.23	5.51	0.787	0.784
Freeform	16	7.34	5.69	0.769	0.750
	32	7.19	5.55	0.783	0.773
	64	7.63	5.91	0.754	0.741
	128	7.62	6.21	0.750	0.733
Northwind	16	7.15	5.33	0.789	0.782
	32	7.09	5.35	0.794	0.788
	64	7.33	5.50	0.785	0.779
	128	7.60	5.81	0.761	0.749

E.3 PREDEFINED LENGTH CANDIDATES

Table 10: Experiments of VAT under different pre-defined segment length candidates.

Lengths	RMSE	MAE	PCC	CCC
2,3,4	7.08	5.25	0.800	0.795
2,3,4,6	7.13	5.13	0.796	0.788
2,3,4,6,8	7.19	5.37	0.793	0.792
2,3,4,6,8,12	6.88	5.17	0.804	0.800
2,3,4,6,8,12,16	7.36	5.47	0.771	0.734

As described in FBSP of the main paper, we pre-define a set of segment length candidates $\mathcal{B}$ (in the unit of FBUs) and predict the segment length d_t for each centre FBU by weighted combination over $\mathcal{B}$. Table 10 examines sensitivity to $\mathcal{B}$ around the shared default $\{2, 3, 4, 6, 8, 12\}$ on AVEC 2013 using VAT, with all other settings fixed. A compact candidate set $(2, 3, 4)$ already yields competitive performance (RMSE/MAE 7.08/5.25), and gradually expanding it to include medium-length segments gives mixed results: $(2, 3, 4, 6)$ slightly improves MAE but not RMSE, while $(2, 3, 4, 6, 8)$ performs worse. The best overall result is achieved by $(2, 3, 4, 6, 8, 12)$, with RMSE/MAE/PCC/CCC of 6.88/5.17/0.804/0.800, suggesting that this range provides sufficient temporal flexibility for modelling depression-related facial behaviours. However, further extending the candidate set to include a very long segment length, i.e., $(2, 3, 4, 6, 8, 12, 16)$, degrades performance to RMSE/MAE 7.36/5.47. This indicates that overly broad length candidates may introduce less informative context and make segment-length prediction harder.

F RELATION TO PRIOR WEAKLY SUPERVISED SEGMENT SELECTION

DSC adopts the category-guided segment ranking of Liao et al. (2026), whose two-branch classification and selection mechanism follows WSDDN (Bilen & Vedaldi, 2016). DMVS differs from Liao et al. (2026) in two main respects. First, the Liao et al.'s method follows a two-stage training pipeline, first training a segment-level encoder and then using its extracted features for segment

selection and video-level loss computation. The second-stage video-level objective therefore learns selection and prediction on fixed backbone features, without updating the encoder. DMVS computes features from the current backbone and propagates the video-level loss back to it through GMVB, enabling end-to-end video-level training. Second, their method constructs segments by concatenating consecutive feature units at predefined temporal scales, without predicting a segment duration conditioned on the local features. FBSP instead predicts a continuous window width at each FBU from a learned distribution over width candidates (Eq.3) and aggregates the features within the resulting boundaries. This enables adaptive variable-length segment construction, allowing segment durations to vary with the input facial behaviour.

G Duration learning through hard segment boundaries

Rounding the FBSP boundaries to integer FBU indices disconnects segment aggregation from the predicted duration d_t. We retain the hard segment representation $\mathbf{s}_t$ defined in the main text and introduce an auxiliary soft aggregation solely to estimate its duration gradient.

Following Sec. 3.2, FBSP constructs one proposal per FBU, giving T proposals on the original FBU feature sequence. For proposal t, the fixed centre is $c_t = t$, and its unrounded edges are $a_t = t - d_t/2$ and $b_t = t + d_t/2$. At FBU-centre coordinate $x_i = i$, we define

$$m_{ti} = \sigma\left(\frac{x_i - a_t}{\epsilon}\right)\sigma\left(\frac{b_t - x_i}{\epsilon}\right), \qquad \epsilon = 2 \text{ FBUs}, \tag{17}$$

where σ denotes the sigmoid function. The soft edges are not clipped, and sums below run over the available T FBUs. Let q_i be the attentive score used by the original aggregation and $w_{ti} = m_{ti}/\sum_j m_{tj}$. We compute

$$\begin{aligned} \mathbf{a}_t &= \sum_i \frac{m_{ti}\exp(q_i)}{\sum_j m_{tj}\exp(q_j)}\mathbf{f}_i, \qquad \boldsymbol{\mu}_t = \sum_i w_{ti}\mathbf{f}_i, \\ [\mathbf{p}_t]_k &= \rho \log\left(\sum_i w_{ti}\exp([\mathbf{f}_i]_k/\rho)\right), \qquad \rho = 1. \end{aligned} \tag{18}$$

Here, $\rho > 0$ is the pooling temperature, k indexes feature channels, and $\mathbf{p}_t$ is a smooth approximation to maximum pooling. Applying the original concatenation and projection to $\mathbf{a}_t$, $\boldsymbol{\mu}_t$ and $\mathbf{p}_t$ gives $\mathbf{s}_t^{\text{soft}}$. We evaluate the normalisations and log-sum-exp in numerically stable form.

During training, we pass the following representation to DSC and VLC,

$$\widetilde{\mathbf{s}}_t = \mathbf{s}_t + \big(\mathbf{s}_t^{\text{soft}} - \text{sg}(\mathbf{s}_t^{\text{soft}})\big), \tag{19}$$

where sg preserves its input value and stops its gradient. The parenthesised term is zero in the forward pass. On the auxiliary path, we hold the centre, FBU features, attentive scores and projection parameters constant, leaving only the duration-dependent weights differentiable. We also detach backbone features at the input to the length branch, while keeping the duration predictor MLP_{len} trainable. The original hard aggregation retains its feature and parameter gradients.

Consequently, the added backward path supplies

$$\widehat{\frac{\partial \mathcal{L}_{\text{vid}}}{\partial d_t}} = \left(\frac{\partial \mathbf{s}_t^{\text{soft}}}{\partial d_t}\right)^{\top} \frac{\partial \mathcal{L}_{\text{vid}}}{\partial \widetilde{\mathbf{s}}_t}. \tag{20}$$

This allows the video-level losses, including the MIL classification loss, to train the length branch without adding gradients to the backbone or aggregation parameters. It is a surrogate gradient, not an exact derivative of integer indexing, and need not be nonzero for every proposal. Segment selection, coverage and GMVB weights retain their hard-window computation. Inference omits the auxiliary aggregation. Thus, at fixed parameters the forward predictions remain unchanged, although subsequent length updates can change them.

H Classification and ordinal supervision

We use two complementary category objectives. The MIL classification loss supervises the category scores aggregated over segment representations, while the ordinal loss supervises ordered severity

thresholds predicted from the video descriptor $\mathbf{z}$. For the four-category depression setting, we index categories by $c \in \{0, 1, 2, 3\}$ and obtain the target from the severity score y as

$$c = \sum_{j=1}^{3} \mathbf{1}_{\{y \geq \beta_j\}}, \qquad (\beta_1, \beta_2, \beta_3) = (14, 20, 29). \tag{21}$$

Thus, equality at a threshold assigns the higher category. We set the coefficients of the regression, ordinal and MIL classification losses to $\lambda_{\text{reg}} = \lambda_{\text{ord}} = \lambda_{\text{cls}} = 1$. The equations below define the per-video classification and ordinal objectives.

MIL classification loss: Let $a_j = \sum_t \mathcal{P}_{t,j}$ be the MIL score for category j, using the joint segment scores defined in DSC. Each a_j is a weighted average of $\hat{\mathcal{H}}_{1,t,j} \in [0, 1]$ with weights $\hat{\mathcal{H}}_{2,t,j}$ that sum to one over segments. Thus, $a_j \in [0, 1]$, but $\sum_j a_j$ need not equal one. The implementation treats these bounded scores as logits for cross entropy, which normalises them across categories, giving

$$p_j = \frac{\exp(a_j)}{\sum_{r=0}^{C-1} \exp(a_r)}, \qquad \mathcal{L}_{\text{cls}}(\mathbf{a}, c) = -\log p_c = -a_c + \log \sum_{r=0}^{C-1} \exp(a_r). \tag{22}$$

Here, $\mathbf{p}$ is a normalised category distribution, whereas $\mathbf{a}$ contains unnormalised MIL scores. This objective differs from both $-\log a_c$ and $-\log(a_c / \sum_j a_j)$. For $C = 4$, the bounded scores imply $p_c \leq e/(e + 3) \approx 0.475$ and $\mathcal{L}_{\text{cls}} \geq \log(e + 3) - 1 \approx 0.744$. We use this loss as auxiliary category supervision and do not interpret $\mathbf{p}$ as calibrated confidence. The score gradient is $\partial \mathcal{L}_{\text{cls}} / \partial a_j = p_j - \mathbf{1}_{\{j=c\}}$, encouraging a higher target-category score and lower competing-category scores.

Category-guided segment selection: DSC first predicts $\hat{c} = \arg\max_j a_j$, then ranks segments using $\mathcal{P}_{t,\hat{c}}$ and selects the top-K to construct $\mathbf{z}$. We use this procedure during both training and inference because selecting with the true category during training would provide label information unavailable at inference. During training, the true category c instead defines the MIL loss $-\log p_c$, using category scores aggregated over all candidate segments. If $\hat{c} \neq c$, this loss still encourages a higher score for c and lower scores for competing categories, including $\hat{c}$. The current forward pass nevertheless uses the segments selected under $\hat{c}$. Gradients are not propagated through the discrete category and segment selections. Instead, they update the continuous scores and representations, which can change the selections in subsequent forward passes.

Ordinal loss: The ordinal head outputs $C - 1$ logits $\hat{q}_k$, with $k \in \{0, \ldots, C - 2\}$. Each sigmoid output $o_k = \sigma(\hat{q}_k)$ estimates whether the target category exceeds threshold index k. We encode the target as $q_k = \mathbf{1}_{\{c>k\}}$ and minimise binary cross entropy averaged over thresholds,

$$\mathcal{L}_{\text{ord}}(\hat{\mathbf{q}}, \mathbf{q}) = -\frac{1}{C-1} \sum_{k=0}^{C-2} \left[q_k \log o_k + (1 - q_k) \log(1 - o_k) \right]. \tag{23}$$

For $C = 4$, categories $0, 1, 2, 3$ correspond to targets $(0, 0, 0)$, $(1, 0, 0)$, $(1, 1, 0)$ and $(1, 1, 1)$, respectively. We compute this expression directly from logits for numerical stability. These are cumulative threshold predictions rather than a categorical softmax distribution. The separate output logits do not impose a hard monotonicity constraint. Unlike MIL classification over all segment scores, the ordinal objective acts on the video descriptor formed from selected segments.

I ACCUMULATIVE GRADIENT COMPUTATION FORMULATION

We distinguish the depression relevance-weighted model update in GMVB from the gradient of the video-level loss. The derivation first gives the unmodulated gradient, then shows how established feature-gradient caching and sequential back-propagation compute it (Gao et al., 2021), and finally states how depression relevance changes each FBU's contribution in model learning.

Notations: Given a video split into T FBUs $\{u_t\}_{t=1}^{T}$, let the target ADA model (backbone) be $\mathcal{M}(\cdot;\theta)$, and denote the per-FBU output feature as

$$\mathbf{f}_t(\theta) = \mathcal{M}(u_t;\theta), \quad t = 1,\dots,T. \tag{24}$$

Let $\mathbf{F}(\theta) = [\mathbf{f}_1(\theta),\dots,\mathbf{f}_T(\theta)]$ denote the FBU feature sequence. The FVC phase (including FBSP/DSC/VLC) induces a scalar loss function $\ell_{\text{vid}}(\cdot,\cdot)$ over the feature sequence and the video label y:

$$\mathcal{L}_{\text{vid}}(\theta) \triangleq \ell_{\text{vid}}\big(\mathbf{F}(\theta), y\big). \tag{25}$$

Note that $\mathcal{L}_{\text{vid}}$ depends on θ only through the set of features $\mathbf{F}(\theta)$.

Full-video back-propagation (infinite-memory reference): Define the FVC gradient w.r.t. each FBU feature as

$$\mathbf{g}_t^{\mathbf{f}} \triangleq \frac{\partial \mathcal{L}_{\text{vid}}}{\partial \mathbf{f}_t}, \quad t = 1,\dots,T. \tag{26}$$

If we back-propagate through the whole video in one pass (i.e., keeping all intermediate features of $\mathcal{M}$), the gradient of θ is given by the multivariate chain rule:

$$\nabla_\theta \mathcal{L}_{\text{vid}}(\theta) = \sum_{t=1}^{T} \left(\frac{\partial \mathbf{f}_t(\theta)}{\partial \theta}\right)^{\top} \frac{\partial \mathcal{L}_{\text{vid}}}{\partial \mathbf{f}_t} = \sum_{t=1}^{T} \left(\frac{\partial \mathbf{f}_t(\theta)}{\partial \theta}\right)^{\top} \mathbf{g}_t^{\mathbf{f}}. \tag{27}$$

Sequential computation of the unmodulated gradient: Our DMVS first computes $\{\mathbf{g}_t^{\mathbf{f}}\}_{t=1}^{T}$ by back-propagating $\mathcal{L}_{\text{vid}}$ through the lightweight FVC blocks (with $\{\mathbf{f}_t\}_{t=1}^{T}$ retained). Then, for each FBU u_t, we re-forward it through $\mathcal{M}$ to reconstruct the required intermediate features $\mathbf{h}_t$ and obtain the per-FBU parameter gradient

$$\mathbf{g}_t^{(\theta)} \triangleq \left(\frac{\partial \mathbf{f}_t(\theta)}{\partial \theta}\right)^{\top} \mathbf{g}_t^{\mathbf{f}}, \tag{28}$$

which is the unmodulated gradient of Eq. 11. With θ unchanged within one video-level step and identical per-FBU forward computations, including any stochastic choices and model state, recomputation gives the same intermediate features and Jacobians. Accumulating these per-FBU gradients over the whole video gives

$$\sum_{t=1}^{T} \mathbf{g}_t^{(\theta)} = \nabla_\theta \mathcal{L}_{\text{vid}}(\theta). \tag{29}$$

Comparing Eq. 27 and Eq. 29, sequential computation gives the same unmodulated gradient $\nabla_\theta \mathcal{L}_{\text{vid}}$ as full-video back-propagation under these assumptions. The same per-FBU chain rule applies to the modulated gradient $[\mathbf{u}]_t\, \mathbf{g}_t^{\mathbf{f}}$ in Eq. 11, allowing the relevance-weighted update below to be computed sequentially.

Effect of the gradient modulation: GMVB back-propagates $[\mathbf{u}]_t\, \mathbf{g}_t^{\mathbf{f}}$ rather than $\mathbf{g}_t^{\mathbf{f}}$, so the update actually applied to the backbone is

$$\sum_{t=1}^{T} \tilde{\mathbf{g}}_t^{(\theta)} = \sum_{t=1}^{T} [\mathbf{u}]_t\, \mathbf{g}_t^{(\theta)}, \tag{30}$$

which equals $\nabla_\theta \mathcal{L}_{\text{vid}}$ when $[\mathbf{u}]_t \equiv 1$. In general, GMVB changes the model update by weighting each FBU's contribution to the video-level loss according to the depression relevance of the selected facial behaviour segments containing it. The scores act as multipliers on the already-computed feature gradients. Sequential computation preserves this weighted update, which differs from the unmodulated loss gradient.

J TRAINING AND MEMORY EFFICIENCY

Training efficiency: We report the training and inference efficiency of DMVS with VAT and LMT backbones on the AVEC 2014 Northwind dataset using an NVIDIA A100 GPU. Table 11 compares the original segment-level training with the corresponding DMVS-augmented method. "Training

Table 11: Training and inference efficiency on AVEC 2014 Northwind dataset.

Method	Param	Training epochs	Per epoch time	Total training time	Inference time	RMSE
VAT	85.63 M	50	144 s	2.0 h	2.29 s	7.69
VAT + DMVS	89.85 M	40	168 s	1.9 h	2.48 s	7.09
LMT	0.96 M	50	101 s	1.4 h	1.81 s	7.84
LMT + DMVS	1.02 M	40	114 s	1.3 h	1.89 s	7.54

epochs" denotes the number of epochs required for convergence, and "Inference time" is the wall-clock time for processing a 1-minute video at 30 FPS.

Compared with segment-level training, DMVS converges in fewer epochs because it optimises video-level supervision directly and avoids the noisy supervision induced by assigning the same video label to every fixed-length segment. Although DMVS introduces additional FVC and GMVB procedures during training, the per-epoch wall-clock time increases only modestly. This is because DMVS performs one optimisation step per video rather than one step per segment, resulting in far fewer optimiser updates per epoch. Consequently, the total training time remains comparable to, or slightly lower than, the original training strategy for both backbones. At inference time, DMVS adds only a small overhead. The GMVB phase is not used during inference, and the remaining extra computation comes from the lightweight MLPs in FVC, which is minor compared with the backbone forward pass.

Memory efficiency: We further analyse the peak activation memory of DMVS when training on long videos. Let a video be partitioned into T non-overlapped FBUs $\{u_t\}_{t=1}^{T}$, and let the backbone $\mathcal{M}(\cdot;\theta)$ produce the per-FBU representation $\mathbf{f}_t = \mathcal{M}(u_t;\theta)$. In conventional full-video end-to-end training, the intermediate activations of the backbone, denoted as $\mathbf{h}_t = \mathcal{M}_{\text{int}}(u_t;\theta)$, must be retained for all T FBUs until the video-level loss is computed and back-propagated. The resulting peak backbone activation memory therefore scales approximately linearly with video length:

$$M_{\text{act}}^{(\text{video})} \approx T \cdot M_{\text{act}}^{\text{FBU}}(\mathcal{M}), \tag{31}$$

where $M_{\text{act}}^{\text{FBU}}(\mathcal{M})$ denotes the activation memory required to process one FBU through $\mathcal{M}$.

In contrast, DMVS avoids storing the backbone activations for all FBUs simultaneously. During FVC, FBUs are forwarded sequentially through $\mathcal{M}$; after each forward pass, $\mathbf{h}_t$ is discarded and only the compact feature $\mathbf{f}_t$ is retained. After the video-level loss $\mathcal{L}_{\text{vid}}$ is computed, GMVB recomputes each FBU forward pass, reconstructs the corresponding $\mathbf{h}_t$, and immediately back-propagates the relevance-weighted feature gradient $[\mathbf{u}]_t\,\mathbf{g}_t^{\mathbf{f}}$, where $\mathbf{g}_t^{\mathbf{f}} = \partial\mathcal{L}_{\text{vid}}/\partial\mathbf{f}_t$. Thus, the backbone activations need to be stored for only one FBU at a time, and the peak activation memory becomes

$$M_{\text{act}}^{(\text{DMVS})} \approx M_{\text{act}}^{\text{FBU}}(\mathcal{M}) + M_{\text{act}}^{\text{seq}}(T, D), \tag{32}$$

where D is the feature dimension. The sequence-level term includes the cached features, their gradients and the FVC activations. With fixed network widths and maximum hard-window length, the feature and segment representations require $O(TD)$ storage. The auxiliary soft-window path additionally retains proposal-by-FBU weights of size $T\times T$, with one row for each of the T proposals and one column for each FBU. Thus, the current implementation has

$$M_{\text{act}}^{\text{seq}}(T, D) \approx 4\big(c_1 TD + c_2 T^2\big) \quad \text{bytes}, \tag{33}$$

in FP32, with storage complexity $O(TD + T^2)$. Here, c_1 and c_2 account for retained linear and proposal-by-FBU quantities, respectively. Features and their cached gradients alone contribute $8TD$ bytes. Soft aggregation uses matrix products without constructing a $T \times T \times D$ tensor. The quadratic term is small over the measured Northwind lengths, as supported by the soft-window overhead reported below. When the complete sequence-level term remains small relative to the backbone activation term, the activation-memory reduction is approximately

$$\frac{M_{\text{act}}^{(\text{video})}}{M_{\text{act}}^{(\text{DMVS})}} \approx T. \tag{34}$$

This approximation is conditional on that regime and does not imply linear memory scaling for arbitrarily long videos.

As a concrete example, consider a 3200-frame video (approximately 100 seconds) with 32 frames per FBU, which gives $T = \lceil 3200/32 \rceil = 100$. With feature dimension $D = 1024$, storing per-FBU features and their gradients requires only

$$M_{\text{features+gradients}}(100, 1024) = 8 \times 100 \times 1024 \text{ bytes} \approx 0.78 \text{ MiB}. \tag{35}$$

For the illustrative activation-only calculation, we use an estimate of 566 MB for one VAT FBU with input $u_t \in \mathbb{R}^{32\times3\times224\times224}$,

$$M_{\text{act}}^{\text{FBU}}(\mathcal{M}(u_t)) \approx 566\text{MB (FP32)}. \tag{36}$$

This is an analytical estimate, not a measurement of all tensors retained by automatic differentiation or of complete training-step memory. The feature and gradient cache costs 8 KiB per FBU at $D = 1024$, excluding FVC activations and the quadratic soft-window term. For a 50-minute AVEC 2013 video with approximately 2,800 FBUs, this cache alone occupies 21.9 MiB. Retaining all backbone activations would require about 1.5 TB under the same activation-only estimate. The complete DMVS peak at this length also depends on the FVC terms and is not established by the shorter-video measurements below. Sequential computation preserves the chosen relevance-weighted update under the assumptions in Appendix Sec. I.

To measure the complete training-step memory, we profiled VAT on the three longest videos in the AVEC 2014 Northwind training/development split. For each video, we measured prefixes of 4, 8, 16, 24, 32, 48 and up to 64–84 FBUs, with two repetitions per available length. The benchmark used FP32, one 32-frame FBU per backbone forward, all configured losses and an AdamW update. Figure 4 reports peak PyTorch-allocated GPU memory and the FVC peak above its phase-entry allocation. The full-backpropagation run reached 38.3 GiB at 32 FBUs and ran out of memory at 48 FBUs on the 40 GiB A100. Checkpointing, GradCache and DMVS completed at 84 FBUs with peaks near 7.17 GiB. At 84 FBUs, the soft-window DMVS used 48.0 MiB above the FVC phase-entry allocation, compared with 46.0 MiB for the identical hard-window DMVS. These measurements include the soft-window masks, aggregation, backward graph and operator temporaries. The 2.0 MiB difference in FVC peaks at 84 FBUs is less than 0.03% of the approximately 7.17 GiB complete training peak. Thus, the additional soft-window storage, including its quadratic component, is small in this measured range. This difference measures the overall auxiliary-path overhead, rather than isolating the quadratic term.

Activation estimates and measured memory: The 566 MB estimate in Eq. 36 concerns backbone activations, whereas Fig. 4 measures complete training peaks, including tensors saved for backward, optimiser states, gradients and temporary allocations. The approximate T-fold activation-memory reduction applies only when the sequence-level terms are small relative to the backbone activations. Figure 4 reports practical training-memory requirements over the measured Northwind lengths.

K VISUALISATION

Fig. 5 shows the relationship between the predicted depression scores and the ground-truth BDI-II scores on the test set, where the predicted scores are produced by the VAT model trained with DMVS. Each point corresponds to one test video, and its distance from the diagonal trend reflects the prediction error for that subject. The predictions are concentrated around the diagonal trend rather than being widely dispersed, indicating that the model preserves the ordering and magnitude of depression severity across subjects. This agreement is consistent with the quantitative RMSE, MAE, PCC, and CCC results, and suggests that the proposed video-level supervision improves not only average error but also the reliability of subject-level severity estimation.

Fig. 6 provides a visual interpretation of how DMVS identifies depression-informative segments. The heat maps visualise the per-FBU relevance scores $\mathbf{u}$ produced in the FVC phase, where warmer colours indicate FBUs that contribute more strongly to the video-level depression prediction. The overlaid coloured bold lines mark the top-10 variable-length facial behaviour segments selected by the model. These selected segments are not uniformly distributed over time and do not share a fixed duration; instead, they concentrate on temporally localised intervals with high relevance scores while adapting their lengths to the duration of the underlying facial behaviours. This supports the motivation of DMVS: depression-related facial cues may appear sparsely and at multiple temporal

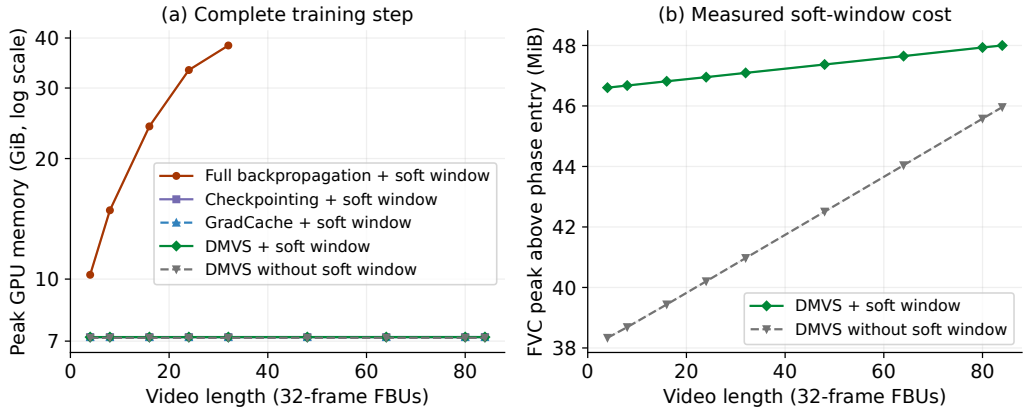


Figure 4: Measured VAT training memory on AVEC 2014 Northwind videos. Left: peak allocated memory across feature extraction, video-level loss computation, backbone back-propagation and the optimiser update. Full back-propagation reaches 38.3 GiB at 32 FBUs and runs out of memory at 48 FBUs. Checkpointing, GradCache and DMVS complete up to 84 FBUs at approximately 7.17 GiB. Right: FVC peak memory above phase entry for DMVS with the soft-window duration-gradient path and the otherwise identical hard-window control. At 84 FBUs the measured increments are 48.0 MiB and 46.0 MiB, respectively. Points average two runs per video and length where available across the three longest videos; the 80- and 84-FBU means contain two and one videos, respectively. Memory denotes PyTorch allocated memory and excludes CUDA context and non-PyTorch allocations.

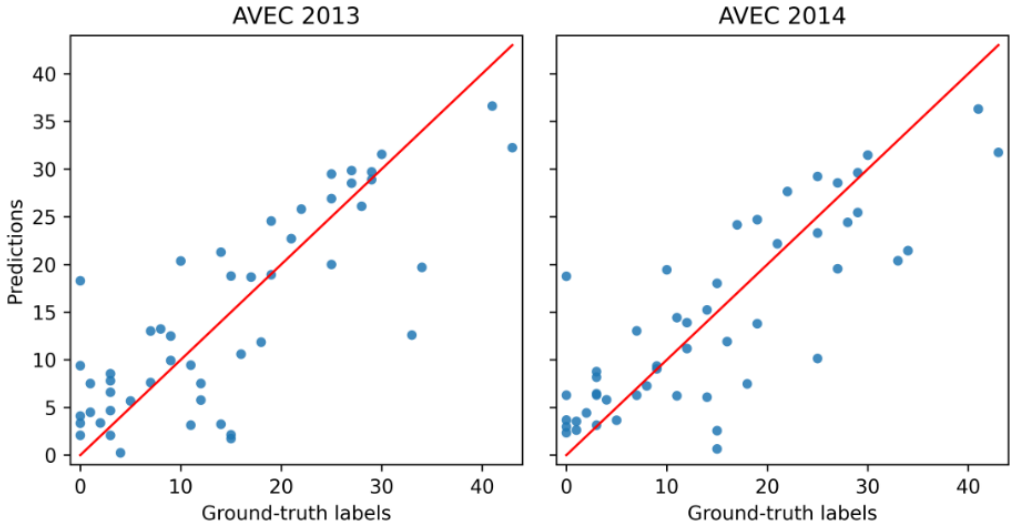


Figure 5: Scatter plot of predicted vs. ground-truth depression scores on the test set.

scales, and video-level supervision enables the model to prioritise the most informative segments rather than treating all fixed-length segments equally.

AU analysis: To further examine whether the segments selected by DMVS correspond to clinically meaningful facial behaviours, we conduct an action-unit (AU) analysis on 5 randomly sampled videos from the Northwind test set. For each video, we rank all candidate segments by their depression relevance scores and select the top-10 and bottom-10 segments. The n-th top-ranked segment is paired with the n-th bottom-ranked segment from the same video, yielding 50 within-video top-bottom pairs in total. Following prior evidence that facial AUs can reflect depressive behavioural cues (Sugimori & Yamaguchi, 2025; Girard et al., 2014), we use OpenFace 2.0 to extract frame-level AU statistics and compare each AU between the paired top and bottom segments. Differences are computed as top minus bottom and summarised by the mean difference and Cohen's d (CD).

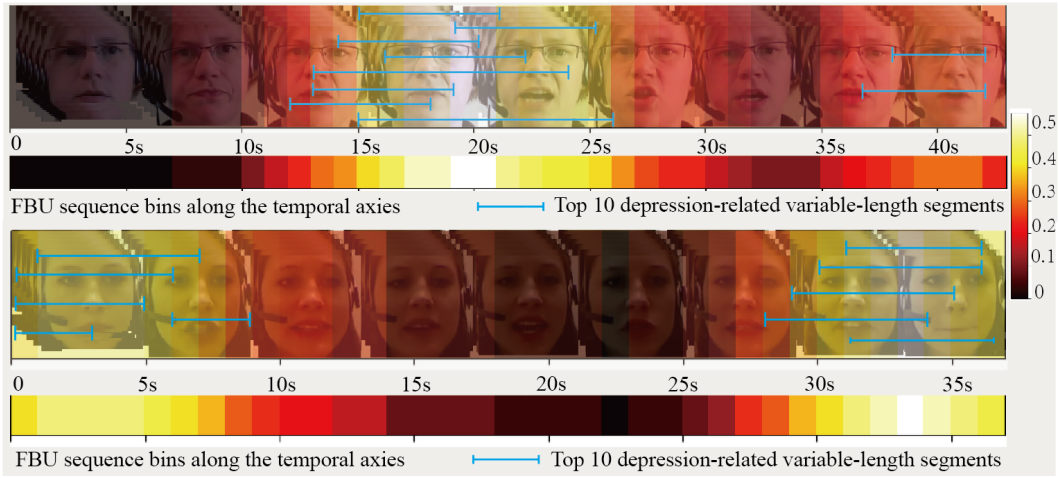


Figure 6: Heat-map visualisation of per-FBU relevance scores from 2 target videos, with the top-10 depression-related segments of variable lengths overlaid above the video frames.

The top-ranked segments show stronger depression-related facial patterns than the bottom-ranked segments. Specifically, AU45 (blink) exhibits higher intensity (+0.039, CD = 0.81) and higher temporal variability (+0.076, CD = 0.70), while AU04 (brow lowerer) also shows increased variability (+0.046, CD = 0.54). In contrast, AU12 (lip corner puller), which is associated with smiling, is reduced in the top-ranked segments (-0.024, CD = -0.19). These results suggest that DMVS assigns higher relevance to segments with more pronounced eye activity, increased brow-related tension, and reduced positive facial expression, aligning with established facial indicators of depressive affect.

L DMVS FOR PERSONALITY ANALYSIS

We further evaluate DMVS on personality analysis using the ChaLearn 2016 personality dataset (Ponce-López et al., 2016), which contains 10,000 15-second video clips. To align with our category-guided segment ranking formulation, we follow Li et al. (2020) and discretise each trait label into four bins, then train the VAT with DMVS for each trait separately. As shown in Table 12, using visual input only, our method achieves the highest mean accuracy and the best results on four of the five traits, with only a small difference on agreeableness, indicating that DMVS can generalise to related video understanding tasks.

Table 12: Results on Chalearn 2016 dataset by percentage. O.,C.,E.,A., N. and M. denote openness, conscientiousness, extraversion, agreeableness, neuroticism and mean respectively.

Method	O.	C.	E.	A.	N.	M.
Zhao et al. (2023)	91.10	91.48	91.30	**91.43**	90.87	91.24
Bohy et al. (2024)	90.80	90.20	89.50	90.70	90.50	90.30
Wang et al. (2025a)	90.90	90.97	90.01	90.53	90.93	90.67
Liao et al. (2024)	91.15	91.23	91.53	90.99	90.98	91.18
Ours	**91.30**	**92.09**	**91.72**	91.27	**91.07**	**91.49**

M LIMITATIONS AND ETHICAL CONSIDERATIONS

Evaluation scope: Our depression experiments are restricted to AVEC 2013 and AVEC 2014. They do not establish robustness across clinical settings, cultures, age groups or acquisition conditions. We do not report demographic subgroup errors, so aggregate improvements cannot establish equitable performance. The targets are self-reported BDI-II severity scores rather than clinician-confirmed diagnoses. Prediction errors therefore combine modelling limitations with uncertainty in

the target measure. The personality experiment evaluates a different construct and does not provide external clinical validation for depression assessment.

Privacy, consent and intended use: Face videos contain identifying information, and inferred mental-health scores are sensitive. Any use beyond the benchmark research setting should require explicit, informed and revocable consent for both recording and mental-health inference, restricted access, data minimisation and a defined retention policy. Consent to recording alone should not be treated as consent to psychological assessment. The method should not be used for covert surveillance or employment, insurance or educational screening. We present DMVS as a research training strategy, not a diagnostic system or a replacement for assessment by qualified clinicians. Any proposed clinical application would require independent validation, appropriate human oversight and a process for contesting or correcting predictions.

Evaluation protocol and interpretation: Our evaluation follows the protocol described in Appendix Sec. A to support comparison with prior results obtained using combined training and development data and test-based checkpoint selection. Replacing these experiments with training on the official training split alone would reduce the training set from 100 to 50 videos and change the basis of comparison. The small sample size and self-reported BDI-II targets also introduce uncertainty into validation-based checkpoint selection. Selecting a training duration on 50 videos and transferring it to training on 100 videos does not guarantee the same optimisation behaviour. Within the stated protocol, Table 3 reports lower RMSE with DMVS than with the corresponding reproduced baselines across five backbones. Table 5 further compares GMVB with unit, mean and shuffled gradient weights over five runs per setting, retaining the same FVC and video-level losses. These results support the effectiveness of DMVS across the evaluated models and the benefit of assigning relevance weights to the corresponding FBUs. This evidence concerns relative performance under the reported protocol. Test-based checkpoint selection can bias both absolute scores and differences between methods. Independent evaluation with a sufficiently large sample would provide more precise estimates of generalisation (Archer et al., 2021; Mao et al., 2023), but access to suitable depression video data remains a shared constraint in the field (Han et al., 2024). Even the larger LMVD dataset withholds raw videos for privacy reasons (He et al., 2026). Larger independent datasets with accessible facial videos and severity labels are therefore a shared priority for evaluating raw-video methods beyond the current ADA benchmark protocol.