# OmniHead: A Unified Model for Dynamic Nonverbal Facial Behaviors

Pierre Vuillecard Jean-Marc Odobez
Idiap Research Institute, Switzerland
Ecole Polytechnique Fédérale de Lausanne, Switzerland
{ pvuillecard, odobez}@idiap.ch

## Abstract

*Nonverbal facial behaviors convey essential cues for human communication, yet most approaches either tackle isolated tasks or focus on unifying static face analysis tasks, overlooking temporal behaviors such as head gestures or gaze shifts. In this paper, we introduce OmniHead, a unified spatiotemporal framework that models dynamic nonverbal face/head behaviors end-to-end. Within a single encoder–decoder multi-task transformer, OmniHead jointly learns gesture, gaze, and affective signals. To overcome the scarcity of labeled data, we propose to leverage unlabeled facial video datasets and perform semi-supervised pretraining via distillation from specialized single-task teachers, which yields an explicit representation of dynamic head behavior. For addressing robust deployment in the wild, we release new annotations on two datasets with head-gesture, blink, and saccade/fixation, enabling as well, for the first time, the study of these behaviors under unconstrained conditions. Experimentally, OmniHead (i) shows the importance of our pretraining approach for the unified model by delivering competitive results compared to task-agnostic pretraining strategies; (ii) achieves state-of-the-art performance on multiple tasks compared to single-task baselines; (iii) exhibits strong cross-dataset generalization. Annotations, code, and models is available at https://github.com/idiap/OmniHead.*

## 1. Introduction

Nonverbal behavior is a fundamental channel of human communication, conveying important information beyond speech, with the face as its richest source of social signals [23]. From this perspective, human interaction is a dynamic and adaptive process in which individuals continuously adjust head gestures, gaze direction, and facial expressions to convey meaning and pursue interpersonal goals. Head gestures reflect attention, agreement, or affective reactions [2, 60, 82]. Gaze direction and behaviors such as blinks, saccades, and fixations are indicative of attentiveness, emotion intensity [38, 48], or mental state [83], while facial expressions reveal intent and emotions [24, 55]. Hence, automatic analysis of facial behavior has received growing interest, enabling diverse applications in human-computer interaction (robotics and driver monitoring) [3, 4, 34, 67, 73], education and entertainment [10, 54, 68], and mental health, where it offers valuable insights into conditions such as depression [26],

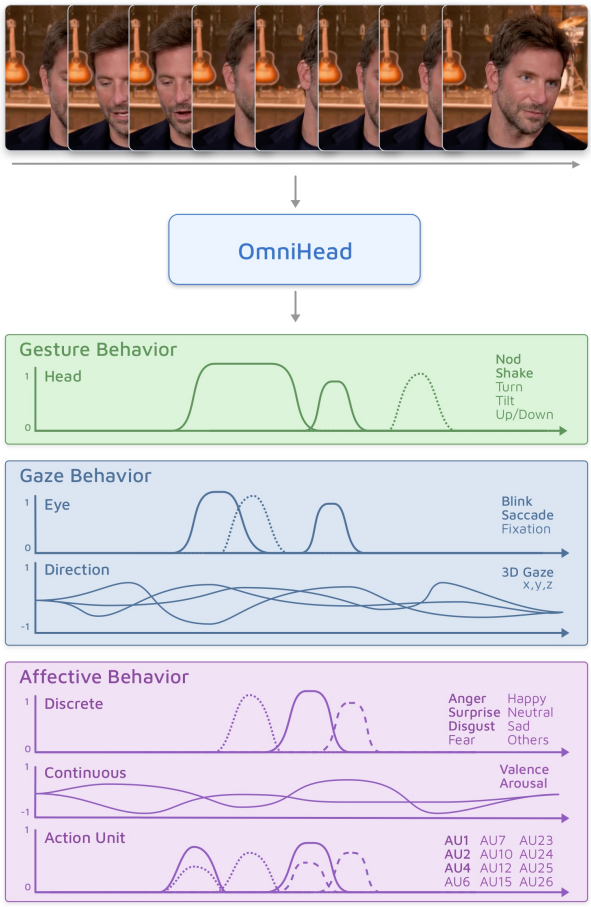



**Figure 1** Previous unified methods focused on static facial analysis. In this work, we propose OmniHead, a unified multi-task spatio-temporal model for dynamic nonverbal facial behaviors including *head gesture*, *blink*, *saccade/fixation*, *3D gaze*, *facial expression*, *valence/arousal*, and *facial action unit* (FAU).

stress [76], and schizophrenia [83]. Recent work has therefore made significant progress in specialized tasks such as head gesture recognition [60, 86], 3D gaze estimation [19, 35, 41, 84, 85, 90], and facial expression recognition [45, 48, 70, 89, 92].

Studying and computing these behaviors in isolation is inefficient for two main reasons: 1) Real-world applications typically require multiple cues simultaneously, leading to increased computational cost for each additional cue; and 2) All tasks are centered on the face, and their behaviors are tightly coupled, shaping facial appearance and dynamics both individually and in combination. It is therefore efficient and natural to model them jointly. To address this, unified approaches have been proposed for facial analysis. HyperFace [66], SwinFace [63], Faceptor [64], and FaceXFormer [56] are single models trained jointly on multiple facial tasks. While effective, these models mainly focus on frame-based perception tasks such as age, race, attributes, face parsing, and landmarks, often overlooking key behavioral cues such as gaze, FAUs, or valence–arousal. OpenFace [8, 9] is one of the few frameworks targeting behavioral analysis, including head pose, gaze, and FAU. However, they are all static methods processing each frame independently, neglecting the natural dynamics and temporal correlations of facial behaviors that often reflect complex cognitive and emotional states. In addition, such methods cannot capture purely temporal behavior, such as head gestures or gaze saccades. Therefore, we introduce a unified model for dynamic nonverbal behavior recognition, encompassing head gestures, 3D gaze, blinks, saccades, FAUs, facial expressions, and valence–arousal, as shown in Fig. 1.

Designing and training a unified model for facial behavior analysis is challenging due to the limited availability and heterogeneity of existing datasets. As collecting large-scale labeled data is labor-intensive, especially for temporally dense annotations and multi-task labels, available datasets are often relatively small and lack diversity. To mitigate this issue, large-scale dataset pre-training shows a speed-up in convergence and helps improve generalization in the small data scenario [64, 95, 97]. For instance, FRL [11] and FaRL [95] adopt task-agnostic pretraining approaches on face images, and MARLIN [13] uses self-supervised learning for facial video representation. However, these pretrained approaches are task-agnostic and do not necessarily capture useful representations for behavior recognition. Therefore, inspired by self-training methods [43, 88], we explore semi-supervised pretraining via specialized model distillation to learn an explicit dynamic head representation on large-scale unlabeled facial video datasets.

Robustness in challenging conditions is essential for real-world deployment of unified models. Yet up to our knowledge, head gestures, blinks, saccades/fixations have not been studied under such conditions. To address this, we annotated two in-the-wild datasets (Video Attention Target (VAT) [20] and ChildPlay [80]) with head-gesture, blink, and saccade/fixation labels, enabling for the first time, end-to-end unified dynamic facial behavior recognition in unconstrained real-world settings.

**Contributions.** To address all these challenges, we introduce OmniHead, the first unified spatiotemporal model for dynamic nonverbal facial behaviors recognition. OmniHead models a wide range of cues, including head gestures, blinks, fixations/saccades, 3D gaze, facial expressions, valence/arousal, and action units, as illustrated in Fig. 1.

Our contributions are as follows:

- **Unified Model.** To the best of our knowledge, OmniHead is the first unified spatiotemporal model for a wide range of dynamic nonverbal facial behaviors.
- **Representation learning.** We leverage unlabeled facial video datasets with semi-supervised learning via distillation of specialized single-task models to learn a strong spatiotemporal behavioral head and face representation. We will release the extracted pseudo-label.
- **New Tasks and Annotations.** Our work addresses for the first time the detection of head gestures, blinks, and saccades/fixations in the wild, and accordingly provides dense annotations on two datasets, Video Attention Target [20] and ChildPlay [80].
- **Competitive Performance.** OmniHead exhibits strong performance in challenging within- and cross-dataset benchmarks, outperforming state-of-the-art single-task models on several tasks.

## 2. Related Works

**Facial Behavior Analysis.** Gaze has been extensively studied, particularly for 3D gaze estimation based on normalized frontal face crops [18, 25, 49, 74, 91, 94]. However, such models fail under extreme head poses. To overcome this limitation, recent works have relied on head crops as input to study "physically unconstrained gaze estimation" [17, 29, 35, 41, 85]. These studies show that, in in-the-wild settings, temporal models can effectively capture head and eye dynamics [17, 29, 30, 35, 58, 85]. Accordingly, we address unconstrained gaze estimation using head crop sequences as input and a spatiotemporal encoder.

Head gesture analysis has remained underexplored, primarily due to the limited availability of annotated datasets. Most of the existing methods rely on sequences of extracted facial cues (e.g., head pose, landmarks, gaze) [60, 61, 86], and typically address the task only for frontal faces. In contrast, we are the first to perform head gesture recognition in an end-to-end manner and under in-the-wild conditions.

Affective behaviors have been widely explored through facial expressions [45, 70, 89, 92], valence–arousal estimation [37], and action units [52]. While these works are focused on static images, other studies investigate dynamic af-

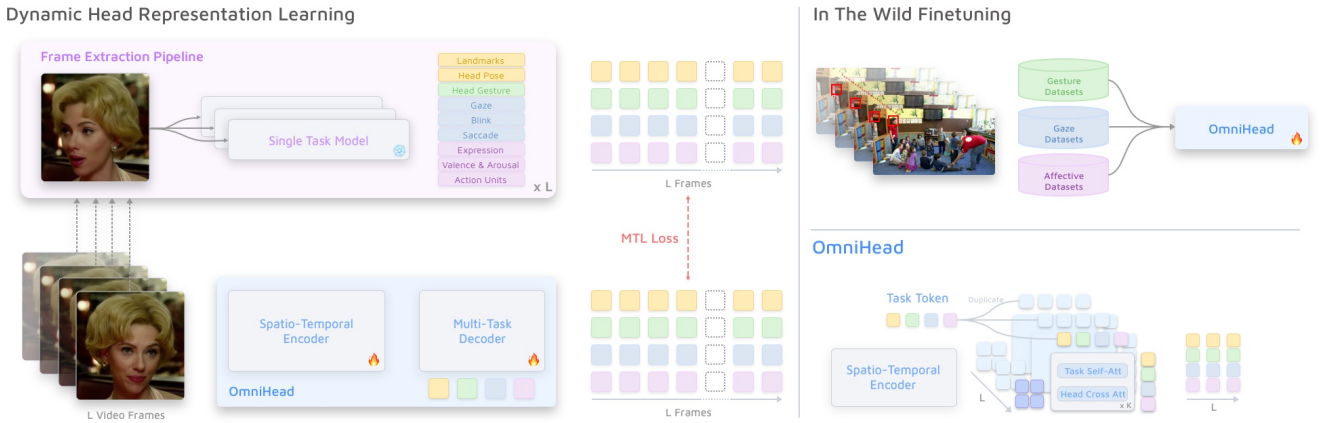



**Figure 2 Approach Overview.** (Left) The goal is to learn a dynamic head representation through explicit facial behavior supervision. We propose a semi-supervised learning framework via specialized models distillation. Specifically, leveraging large-scale unlabeled facial video datasets such as CelebV-HQ [96], we extract pseudo-labeled sequences using expert single-task models on frames. OmniHead is then trained to predict these pseudo-label sequences via a multi-task loss. (Top-right) The pretrained model is further aligned to in-the-wild scenarios using challenging labeled datasets, enhancing robustness in real-world conditions. OmniHead achieves state-of-the-art performance across diverse facial behavior tasks and demonstrates strong generalization. (Bottom-right) Architectural overview of OmniHead, which employs an encoder-decoder transformer with learnable task-tokens.

fective behavior. Notably, Kollias *et al*. [40] introduced the Aff-Wild2 dataset with three affective behaviors densely annotated, and later proposed a multi-task challenge [39] for their joint prediction. [47, 57] show that explicit task sharing via a cascade decoder improved performance, and [72] shows that late temporal smoothing is also effective. Similarly, we used a task-token transformer decoder to explicitly share information across tasks, but we employ a spatiotemporal encoder to model facial dynamics early.

Finally, while some works have addressed head [16] and gaze gesture [75] detection from remote sensors in specific human interactions, we are not aware of works that have addressed the recognition of head gestures, fixations, saccades or blinks in in-the-wild settings.

**Multi-task Approaches.** Multi-task learning has been extensively studied, enabling unified architectures that are both efficient and capable. Recent approaches in the face domain (Faceptor [64], FaceXFormer [56]) introduced unified encoder–decoder transformers relying on task-token transformer decoders inspired by DETR [14]. All existing facial unified methods [63, 66] primarily address static appearance tasks (e.g., parsing, attribute), rather than behavioral tasks besides facial expression recognition. OpenFace framework [8, 9] is one of the few multi-task models that includes gaze and action units. Yet, none of them investigates temporal behaviors such as head gestures, blinks, or saccades/fixation, focusing on static image analysis. Our approach is the first to encompass a broad range of behavioral tasks, including those that are purely temporal, and investigate a spatiotemporal unified model.

**Universal Facial Representation Learning.** Following a general trend, other approaches like FRL [11] and FaRL [95] have explored task-agnostic unsupervised representation learning such as SwAV [15], image–text contrastive loss [65], or masked image reconstruction [31] to learn a unified static face representation, which is further fine-tuned per task on small annotated datasets. Spatiotemporal face models have also been attempted. MARLIN [13] and follow-up works [77, 78] have used various temporal masking autoencoder (e.g. by densely masking salient facial parts like eyes or mouth), similar to VideoMAE [81]. However, temporal facial reconstruction alone struggles to capture subtle inter-dependent nonverbal dynamic behaviors. Furthermore, all the above models (static, dynamic) do not learn multi-task models, only fine-tuning the unsupervised representation on single tasks. To learn generalist face models, a more robust pretraining is crucial as documented in [64]. Therefore, in this work, we propose a semi-supervised pretraining approach to explicitly and jointly learn complex inter-related nonverbal dynamical behaviors.

## 3. Task definitions and Annotations

This work is the first to address dynamic facial nonverbal behaviors using an end-to-end unified model. We present these behaviors, grouped into three categories as in Fig. 1, along with auxiliary geometric tasks.

### 3.1. Tasks

**Gesture Behavior.** Head gestures are important in communication. Following the definition from CCDb-HG [86], we categorize them into six coarse gestures: Nod, Shake, Tilt, Turn, Up/Down, and None. Note that while Nod, Shake, and Tilt are mainly communicative gestures, Turn and Up/Down are gestures more related to attention shifts.

| | *Head Track* | | | | *Head Gesture* | | | | *Blink* | | | | *Saccade* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | | Test | | Train | | Test | | Train | | Test | | Train | | Test | |
| Datasets | # | ≈ | # | ≈ | # | ≈ | # | ≈ | # | ≈ | # | ≈ | # | ≈ | # | ≈ |
| VAT [20] | 949 | 92 | 298 | 82 | 2583 | 12 | 512 | 10 | 833 | 4 | 202 | 5 | 2378 | 7 | 692 | 6 |
| ChildPlay [80] | 734 | 179.5 | 118 | 133 | - | - | 561 | 12 | 644 | 5 | 93 | 4 | 3587 | 7 | 459 | 7 |

**Table 1 New Annotations in-the-wild.** Statistics of the temporal annotation for head gesture, blink, and saccade. # is the number of events and ≈ is the median events duration in frame.

Each frame is annotated with one gesture.

**Gaze Behavior.** Gaze behaviors include direction, blink, fixation, saccade, and smooth pursuit. We exclude smooth pursuit as it is relatively rare and difficult to annotate. Following [85], gaze direction is represented as a 3D unit vector. We defined blink as a binary detection task. As illustrated in Fig. 1, the first frame contains a blink, whereas in the second frame the person looks downward. Although the frames appear visually similar, their temporal dynamics disambiguate the underlying behaviors. Eye movements can be categorized into fixations and saccades. A saccade (or gaze shift) is a rapid eye movement separating one fixation from another [44] (e.g., between frames 3 and 4 in Fig. 1). We formulated this as a two-class problem: fixation vs. shift. Note that saccade dynamics and blink rate correlate with cognitive attentional demand [7].

**Affective Behavior.** Facial expressions are powerful indicators of emotion and have been widely studied through three main tasks. *Facial expressions*, represented as eight discrete emotions: Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Other(or Comptempt in some works [70]). *Valence/arousal*, which are 2D score in the range [-1,1] allowing to provide a continous representation of emotions, with valence measuring the positivity of the emotion, and arousal its intensity [69]. Finally, *facial action units* represent individual facial muscle activations defined by the Facial Action Coding System (FACS) [22] and form the building blocks of complex expressions. Following [52], we restrict them to twelve multilabel classes.

**Head Geometry.** All human heads have similar geometry with slight variations. Geometry gives important knowledge about the head orientation, eyes, and mouth location. Therefore, we include it as an auxiliary task during the pretraining stage. Following the Flame 3D model [46], we encode the head geometry using six parametrized sets of vector: shape $\in \mathbb{R}^{300}$, expression $\in \mathbb{R}^{100}$, pose $\in \mathbb{R}^{6}$, jaw $\in \mathbb{R}^{3}$, translation $\in \mathbb{R}^{3}$, and scale $\in \mathbb{R}$.

### 3.2. Annotations

To address the lack of annotations in-the-wild for gesture, blink, and saccade/fixation recognition, we annotated two in-the-wild datasets, Video Attention Target (VAT) [20] and ChildPlay [80], with frame-level labels for these behaviors. Each head track is annotated and verified by trained annotators. Frames where the behavior is not clearly visible are annotated with $-1$ and excluded from gaze-related training and evaluation. Annotation statistics are provided in Tab. 1. These annotations enable, for the first time, the study of dynamic facial behavior under realistic conditions. More details about the annotation process are presented in Sec. C

## 4. Approach

In this section, we present the components of our proposed framework, whose overview is illustrated in Fig. 2. We first present the OmniHead architecture, an encoder–decoder multi-task framework designed to predict temporally dense facial behaviors. Next, we describe our semi-supervised pretraining strategy, which exploits pseudo-labels from specialized single-task models to learn dynamic behavioral representations from large-scale unlabeled videos. Finally, we detail the in-the-wild alignment stage, where OmniHead is fine-tuned on annotated, challenging datasets to ensure robust generalization across diverse real-world conditions.

### 4.1. OmniHead Architecture

OmniHead is an encoder-decoder multi-task model that predicts temporally dense facial tasks. Formally, given a head crop video clip $\mathbf{X} \in \mathbb{R}^{L\times H\times W\times 3}$, OmniHead output predictions $\mathbf{T}_{i,j}^{out}$ for each frame $i \in [1, ..., L]$ and task $j \in [1, .., N]$ in an end-to-end manner, defined as:

$$\text{OmniHead}(\mathbf{X}) = \{\mathbf{T}_{i,j}^{out}\}_{i=1,j=1}^{L,N} \tag{1}$$

#### 4.1.1. Spatio-Temporal Encoder

**Motivation.** Our framework is encoder-agnostic, yet a unified model must satisfy three key requirements: 1) capturing subtle temporal variations in the input, such as head motion and gaze shifts; 2) extracting fine-grained local signals, like the eye region for gaze-related tasks; and 3) capturing global information such as head orientation. Vision Transformers (ViTs) have shown strong performance in facial analysis, as demonstrated by Faceptor [64], particularly after large-scale pretraining [13, 95]. However, as noted by Cheng *et al*. [19], ViTs are less effective for gaze estimation, as they may overlook critical local eye details. In contrast, hierarchical transformers such as Swin [50, 51] better capture fine-grained local information (e.g., with $4 \times 4$ patch sizes), enabling state-of-the-art results in face analysis like GaT for gaze estimation [85] or FaceXformer for multiple facial tasks [56].

**Tokenization.** The input clip $\mathbf{X}$ is divided into a set of 4D sub-tensors (patches) $\{\mathbf{x}_q\}_{q=1}^{M}$, where $\mathbf{x}_q \in \mathbb{R}^{t\times\hat{h}\times\hat{w}\times 3}$. Following [27, 28, 53, 81], we set $t = 2$, which preserves optical flow and captures subtle short-term motion cues like slight head movements (e.g., nods) or eye motions (e.g., saccades). A linear layer followed by LayerNorm is then applied to project each patch into a token representation.

**Spatio-temporal Encoder.** The patch tokens are then pro-

cessed by a Video Swin hierarchical spatiotemporal encoder [51]. Due to its hierarchical structure, the number of spatial tokens decreases as the network deepens, and the temporal dimension is halved because the patch split uses $t = 2$. To preserve the original temporal resolution (length $L$), we apply an interpolation function to the frame-level feature maps output by the Swin transformer along the temporal axis, resulting in $L$ feature maps $\mathbf{F}_i \in \mathbb{R}^{h \times w \times d}$.

#### 4.1.2. Multi-Task Decoder

Following a DETR-style decoding scheme, we employ a transformer decoder with a set of learnable task tokens $\mathbf{T}^{inp} = \{\mathbf{T}_j^{inp}\}_{j=1}^{N}$, where each $\mathbf{T}_j^{inp} \in \mathbb{R}^d$ represents a specific task. Each task token learns task-specific representations by alternately interacting via self-attention with other task tokens $\mathbf{T}^{inp}$, and cross-attention with frame tokens $\mathbf{F}_i$, thereby enriching the shared representation space. The decoder operates independently on each frame feature map $i$, encouraging the encoder to model and learn the required temporal dynamics. Accordingly, the input task tokens are duplicated for each frame as $\mathbf{T}_{i,j}^{inp} = \mathbf{T}_j^{inp}$. The Multi-Task Decoder consists of $K$ layers, each layer $l$ comprising a Task Self-Attention and a Task-to-Frame Cross-Attention module, as illustrated in Fig. 2. Hence, formally, for each frame $i$, the decoder computes:

$$\{\mathbf{T}_{i,j}^{K}\}_{j=1}^{N} = \text{MultiTaskDecoder}(\mathbf{T}^{inp}, \mathbf{F}_i) \tag{2}$$

**Task Self-Attention.** Facial tasks are often correlated. For example, blinks often occur during saccades. To capture such dependencies, we apply a self-attention mechanism to enable interaction among task tokens $\mathbf{T}^l$. The self-attention is not applied in the first block. The updated task tokens $\tilde{\mathbf{T}}^l$ are computed as: $\tilde{\mathbf{T}}^l = \text{SelfAttention}(\mathbf{T}^l)$

**Task-To-Frame Cross-Attention.** Each facial task depends on features at specific spatial locations and resolutions, e.g., 3D gaze estimation requires detailed eye-region features, while head gesture recognition requires global head orientation variations. To account for these differences, we apply a cross-attention mechanism enabling each task token to selectively attend to and aggregate relevant information from the frame representation according to $\mathbf{T}^{l+1} = \text{CrossAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, with $\mathbf{Q} = \tilde{\mathbf{T}}^l$, $\mathbf{K} = \mathbf{F}_i + \mathbf{E}_{\text{pos}}$, $\mathbf{V} = \mathbf{F}_i$. where $\mathbf{E}_{\text{pos}}$ is a fixed spatial positional embedding.

**Predictions.** After the multitask decoder, each task token $\mathbf{T}_{i,j}^{K}$ is used to predict its task vector output using a task-specific multi-layer perceptron $\text{MLP}_j$ with 2 hidden layers.

$$\mathbf{T}_{i,j}^{out} = \text{MLP}_j(\mathbf{T}_{i,j}^{K}) \tag{3}$$

### 4.2. Dynamic Head Representation

In contrast to FaRL [95] and MARLIN [13], which are task-agnostic self-supervised approaches, our goal is to explicitly learn dynamic behavioral representations. To this end, we adopt a semi-supervised pretraining strategy that leverages pretrained single-task frame models through pseudo-label supervision. By utilizing large-scale, high-quality unlabeled video datasets, we aim to enhance the model's dynamic head representation by supervising it with weak but behaviorally informative facial signals.

#### 4.2.1. Pipeline

As illustrated in Fig. 2, we sample a short head crop video clip $\mathbf{X}$ from a large-scale unlabeled facial video dataset.

**Pseudo-labeling.** For each frame $\mathbf{X}_i$, we apply a pretrained single-task model $\Psi_j$ corresponding to task $j$ to generate a pseudo-label $\mathbf{P}_{i,j} = \Psi_j(\mathbf{X}_i)$.

**Multi-Task Learning Loss.** We supervise each task and frame prediction $\mathbf{T}_{i,j}^{out}$ with the corresponding pseudo-label $\mathbf{P}_{i,j}$ through a task-specific loss $\mathcal{L}_j$. The OmniHead representation is optimized by minimizing the combined objective:

$$\mathcal{L}_{\text{MTL}} = \sum_{j=1}^{N} \sum_{i=1}^{L} w_j \mathcal{L}_j(\mathbf{T}_{i,j}^{out}, \mathbf{P}_{i,j}) \tag{4}$$

Where $w_j$ denotes the loss weight for task $j$. Determining optimal weights is nontrivial and has been extensively studied. We experimented using adaptive weight based on the homoscedastic uncertainty of each task [36], but found it to perform slightly worse than empirically selected weight. More details in appendix G.2.

#### 4.2.2. Pseudo-labels and losses

Details on each pseudo-label and loss are in Appendix E.

**Gesture Behaviors.** *Pseudo-label:* The method 1DCNN [86] relies on Mediapipe[12], but Mediapipe fails to detect faces in challenging head poses. Therefore, we retrained 1DCNN [86] on facial landmarks and head poses extracted using VGGHead [42], improving robustness under non-frontal views. This retrained model is denoted as 1DCNN-Flame. *Task token:* A single task token is assigned to represent head gestures. *Loss:* Supervision uses a standard cross-entropy loss $\mathcal{L}_{gest}$.

**Gaze Behaviors.** *Pseudo-labels.* Robust 3D gaze directions are obtained using ST-WSGE GaT [85]. Since no reliable in-the-wild models exist for blink and saccade detection, these labels are derived using a threshold-based algorithm applied to temporal 3D gaze trajectories. *Task tokens.* One token per gaze task is used. *Losses.* Angular loss $\mathcal{L}_{gaze}$ for 3D gaze regression; standard cross-entropy losses $\mathcal{L}_{blin}$ and $\mathcal{L}_{sacc}$ for blink and saccade prediction, respectively.

**Affective Behaviors.** *Pseudo-labels:* They are derived from specialized pretrained models. Facial expression: [70] trained on AffectNet [57]. Valence/arousal: ELIM [37] trained on AffectNet and Aff-Wild [40]. Action units: GraphAU [52] trained on BP4D [93]. *Task tokens:* One task token per task is used. *Losses:* Following [39], cross-entropy loss $\mathcal{L}_{expr}$ is used for facial expression, binary

cross-entropy $\mathcal{L}_{au}$ for action units, and concordance correlation coefficient loss $\mathcal{L}_{va}$ for valence/arousal.

**Head Geometry.** *Pseudo-labels:* Frame-level 3D head geometry is obtained using VGGHead [42], which estimates FLAME parameters. *Task tokens:* One task token per FLAME parameter. A Flame layer [46] is used after the MLP to transform the FLAME parameters to 3D head vertices, landmarks, and head pose matrix. *Losses:* Following [42], the supervision includes a 2D reprojection loss $\mathcal{L}_{land}$ for landmarks, mean absolute error $\mathcal{L}_{vert}$ for 3D vertices, and geodesic loss $\mathcal{L}_{pose}$ for head pose.

### 4.3. In-the-Wild Alignment

Our pretraining strategy relies on pseudo-labels generated by pretrained single-task models. Consequently, OmniHead must be aligned with ground-truth annotations from in-the-wild datasets to achieve robustness under challenging real-world conditions. We evaluate two fine-tuning setups to assess both task-specific and unified performance. In both cases, we further perform a comprehensive cross-dataset evaluation to quantify the model's generalization capability.

**Single-Task Learning (OmniHead$_{\text{STL}}$).** Here, OmniHead is fine-tuned independently for each task $j$ using its corresponding ground-truth dataset. This allows us to evaluate the pretraining strategy effectiveness on individual tasks and to establish a baseline for comparison with the unified model. During training, only the task-specific loss $\mathcal{L}_j$ is applied.

**Multi-Task Learning (OmniHead$_{\text{MTL}}$).** Here, OmniHead is fine-tuned jointly across all tasks. Because the datasets vary in size and task annotations, a balanced training strategy is required. Following prior unified models [27, 32], we sample a fixed number of instances from each dataset per epoch, undersampling larger datasets and oversampling smaller ones. During training, each batch is randomly drawn from a single dataset, and only the loss terms corresponding to the tasks annotated in that dataset are exploited.

## 5. Experiments

In this work, we use one large-scale video dataset for pretraining and 7 datasets to further train and evaluate OmniHead. We experiment on both STL and MTL settings and compare with STL state-of-the-art (SOTA) methods.

Assessing the generalization capability of OmniHead is crucial for real applications. Therefore, we evaluate our model in two settings:

- *within-dataset*: training and evaluation are performed on the same dataset(s);
- *cross-dataset*: models are evaluated on unseen datasets to assess generalization.

### 5.1. Datasets and Metrics

More information on each dataset, including samples images, is in appendix B.

**Pretraining dataset.** We use CelebV-HQ [96], a large facial video dataset with 35k clips, wide diversity with 15k identities, and a high resolution 512×512 crucial to extract qualitative pseudo-label. Head poses are near frontal ±80°.

**Within-dataset ⟳.** In this evaluation setting (⟳ symbol), we use: *Head Gesture*: CCDb-HG [6, 86]; *Blink Fixation/Saccade*: Video Attention Target (VAT) [20]; *3D Gaze*: Gaze360 [35]; *Expression, Valence/Arousal, Action Unit*: s-Aff-Wild2 [39, 40].

**Cross-datasets ⇆.** In this setting, models are trained on the within-datasets and tested on: *Head Gesture*: KTH [59], VAT [20] ChildPlay [80]; *Blink*, *Fixation/Saccade*: ChildPlay; *3D Gaze*: GFIE [33].

Datasets collected in-the-wild better reflect real-world application conditions. Here, in-the-wild refers to natural video footage involving unconstrained head poses, facial appearance variability, and difficult imaging conditions (e.g., illumination, resolution). CCDb and KTH contain mostly near-frontal faces at fixed sizes and with a fixed viewpoint, whereas VAT, ChildPlay, Gaze360, and GFIE feature unconstrained head poses. s-Aff-Wild2 is also considered an in-the-wild dataset due to its challenging video quality and variations in head pose.

**Evaluation metrics.** OmniHead predictions are computed on every frame using a sliding window of size $L$, keeping the central frame output as the prediction. $F_1$ is selected as the main metric for all classification tasks, following established datasets evaluation protocols for fair comparison. This is a valid choice given the large class imbalance on most tasks. Note that the $F_1$ metric is computed at both frame $F_1^F$ and event $F_1^E$ levels (after detection-groundtruth event association, see [86] repository). Furthermore, both micro (e.g. $F_1^{F_{mi}}$) and macro ($F_1^{F_{ma}}$, average over class results) are computed for multi-class tasks, as an increase in macro better reflects the improvement of minority classes compared to micro. The mean Concordance Correlation Coefficient $CCC$ [39] is used for valence and arousal evaluation. Angular error in degrees is used for 3D gaze estimation [84]. In the result discussion, all reported percentage comparisons are relative values. When grouped by tasks, we report the average.

### 5.2. Implementation Details

Compared to gaze [1, 19, 84] and affective [47, 72, 87] studies that rely on face crops, we use head crops to improve robustness under extreme head poses, at the cost of reduced resolution in frontal views [85], which is known to impact performance. To obtain unified inputs across datasets, we first apply the same robust head detector[1] to each dataset. Each sample consists of a clip of $L = 16$ frames, with the temporal stride adjusted per dataset so that each clip spans approximately one second. Heads are cropped based on the

[1] https://github.com/zhangda1018/yolov5-crowdhuman

| *Gesture Behavior* | | | | | *Gaze Behavior* | | | | | | *Affective Behavior* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CCDb** *Head Gesture* | | | | **VAT** *Blink* | | **VAT** *Saccade* | | **Gaze360** *3D Gaze* | | | **sAff-Wild2** *Expression* | **sAff-Wild2** *Valence-Arousal* | **sAff-Wild2** *Action Units* | – *Total* |
| | $F_1^{F_{mi}}$ | $F_1^{F_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^F$ | $F_1^E$ | $F_1^F$ | $F_1^E$ | | Full | | $F_1^{F_{ma}}$ | $CCC$ | $F_1^{F_{ma}}$ | $\sum$ |
| Paggio [61, 86] | 0.53 | 0.21 | 0.50 | 0.22 | Baseline | 0.22 | 0.22 | Baseline | 0.64 | 0.69 | GaT ST-WSGE† [85] | 12.17 | EfficientNet-B2† [71] | 0.38 | 0.30 | 0.46 | 1.15 |
| Otsuka [60, 86] | 0.52 | 0.41 | 0.53 | 0.45 | | | | | | | Gaze360 [35] | 13.5 | Cross-attention† [57] | 0.33 | 0.50 | 0.47 | 1.30 |
| 1DCNN [86] | 0.68 | 0.51 | 0.69 | 0.56 | | | | | | | Kothari [41] | 13.2 | MT-EmotiDDAMFN† [72] | 0.33 | 0.48 | 0.49 | 1.31 |
| 1DCNN-Flame* | 0.72 | 0.52 | 0.72 | 0.55 | | | | | | | MCGaze [29] | 12.96 | VGGFace [39] | - | - | - | 0.32 |
| | | | | | | | | | | | GaT [85] | 12.60 | CNN-Transformer [87] | - | - | - | 0.98 |
| | | | | | | | | | | | MSA+Seq [17] | 12.50 | Dino-Graph [47] | **0.35** | **0.47** | 0.43 | **1.25** |
| OmniHead$_{\text{STL}}$ | **0.74** | **0.63** | **0.74** | **0.67** | **0.54** | **0.67** | **0.66** | **0.70** | | **11.78** | | | | 0.34 | 0.46 | **0.45** | **1.25** |
| OmniHead$_{\text{MTL}}$ | 0.73 | 0.60 | 0.72 | 0.65 | 0.53 | 0.62 | 0.65 | 0.67 | | 12.50 | | | | 0.30 | 0.44 | 0.44 | 1.18 |

**Table 2 Within-datasets comparison with State-of-the-art.** Within-dataset comparison against state-of-the-art (SOTA) STL methods, with OmniHead evaluated in both STL and MTL settings. Methods† marked with † use additional ground-truth labels for pretraining and are thus less comparable. Methods marked with * are trained by ourselves. **Best** and second best results are highlighted.

| *Gesture Behavior* | | | | | | | | | *Gaze Behavior* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **KTH** *Head Gesture* | | | | **VAT** *Head Gesture* | | **ChildPlay** *Head Gesture* | | **ChildPlay** *Blink* | | **ChildPlay** *Saccade* | | **GFIE** *3D Gaze* |
| | $F_1^{F_{mi}}$ | $F_1^{F_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^F$ | $F_1^E$ | $F_1^F$ | $F_1^E$ | Full |
| Paggio [61, 86] | 0.35 | 0.16 | 0.32 | 0.15 | - | - | - | - | Base. | 0.13 | 0.07 | Base. | 0.50 **0.62** GaT [85] 20.89 |
| Otsuka [60, 86] | 0.47 | 0.31 | 0.47 | 0.35 | - | - | - | - | | | | | |
| 1DCNN [86] | 0.60 | 0.43 | **0.65** | 0.48 | - | - | - | - | | | | | |
| 1DCNN-Flame | **0.63** | 0.45 | 0.64 | 0.49 | 0.53 | 0.35 | 0.32 | 0.24 | | | | | |
| OmniHead$_{\text{STL}}$ | 0.50 | 0.39 | 0.54 | 0.44 | **0.61** | **0.43** | **0.40** | **0.32** | **0.41** | 0.44 | **0.51** | 0.61 | **20.65** |
| OmniHead$_{\text{MTL}}$ | 0.52 | **0.46** | 0.52 | **0.50** | 0.60 | 0.34 | **0.40** | 0.31 | **0.41** | **0.48** | **0.51** | 0.59 | 21.44 |

**Table 3 Cross-datasets comparison with State-of-the-art.** Comparison with SOTA methods in both STL and MTL training settings. Cross-datasets are not used during training except for VAT where OmniHead$_{\text{MTL}}$ has been trained on it for *blink* and *saccade*.

| | *Gesture* | | | | | | *Gaze* | | | | | | *Affective* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CCDb** ↻ *Head Gesture* | | **KTH** ⇆ *Head Gesture* | | **VAT** ⇆ *Head Gesture* | | **VAT** ↻ *Blink* | **ChildPlay** ⇆ *Blink* | **VAT** ↻ *Saccade* | **ChildPlay** ⇆ *Saccade* | **Gaze360** ↻ *3D Gaze* | **GFIE** ⇆ *3D Gaze* | **sAff-Wild2** ↻ *Expression* | **sAff-Wild2** ↻ *Valence-Arousal* | **sAff-Wild2** ↻ *Action Units* |
| | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^E$ | $F_1^E$ | $F_1^E$ | $F_1^E$ | Full | Full | $F_1^{F_{ma}}$ | $CCC$ | $F_1^{F_{ma}}$ |
| STL | | | | | | | | | | | | | | | |
| MARLIN [13] | 0.69 | 0.52 | **0.60** | 0.41 | 0.56 | 0.32 | 0.37 | 0.19 | 0.63 | 0.55 | 12.63 | 24.23 | 0.13 | 0.16 | 0.33 |
| VideoMAE [81] | 0.69 | 0.61 | 0.54 | **0.47** | **0.63** | **0.43** | 0.59 | 0.38 | 0.69 | **0.64** | 12.39 | 22.99 | 0.18 | 0.29 | 0.40 |
| OmniHead$_{\text{STL}}$ | **0.74** | **0.67** | 0.54 | 0.44 | 0.61 | **0.43** | **0.67** | **0.44** | **0.70** | 0.61 | **11.78** | **20.65** | **0.34** | **0.46** | **0.45** |
| MTL | | | | | | | | | | | | | | | |
| MARLIN [13] | 0.69 | 0.61 | 0.39 | 0.24 | 0.45 | 0.15 | 0.55 | 0.30 | 0.63 | 0.54 | 13.37 | 26.53 | 0.21 | 0.27 | 0.38 |
| VideoMAE [81] | **0.73** | 0.64 | 0.32 | 0.17 | 0.54 | 0.20 | 0.57 | 0.30 | 0.63 | 0.55 | 13.12 | 25.07 | 0.19 | 0.24 | 0.37 |
| OmniHead$_{\text{MTL}}$ | 0.72 | **0.65** | **0.52** | **0.50** | **0.60** | **0.34** | **0.62** | **0.48** | **0.67** | **0.59** | **12.50** | **21.44** | **0.30** | **0.44** | **0.44** |

**Table 4 Pretraining.** Comparison of our semi-supervised pretraining strategy against SOTA self-supervised video models. ↻ within and ⇆ cross datasets evaluation.

central frame bounding box and resized to $224 \times 224$, a crucial step to preserve subtle motion cues in the input.

All models used for training have comparable size (≈28M parameters): VideoSwin-T (OmniHead), ViT-S (MARLIN, VideoMAE). Training details are provided in Appendix F.3.

## 5.3. Comparison with the State-of-the-art (SOTA)

Comparisons with SOTA are reported in Tab. 2 (within-dataset), Tab. 3 (cross-dataset), and Tab. 4 (pretraining). Note that the blink, saccade, and 1DCNN-Flame baselines are identical to those used for pseudo-label extraction in the pretraining pipeline (Sec. 4.2.2 and Appendix D). Overall, compared to SOTA, OmniHead achieves competitive performance and yields notable gains for underrepresented classes such as tilt and blink. It further demonstrates strong generalization to in-the-wild datasets, particularly for head gesture and blink recognition.

**Within-datasets.** Tab. 2 shows that single-task OmniHead$_{\text{STL}}$ models match or improve over SOTA. More specifically, we have: (i) a gain in *Gesture*, especially on minority classes, with an 18% boost in macro event $F_1^{E_{ma}}$; (ii) 46% improvement on blink compared to the baseline, which is mainly suited for near-frontal faces and struggles with VAT's head poses; (iii) 6% improvement in 3D gaze estimation; and (iv) marginal improvement on saccade. In *Affective*, OmniHead$_{\text{STL}}$ perform on par with SOTA.

Despite the challenges of having a single unified model for facial behavior prediction, OmniHead$_{\text{MTL}}$ only exhibits a marginal drop of performance compared to all OmniHead$_{\text{STL}}$ models. With respect to SOTA, OmniHead$_{\text{MTL}}$ remains the best on *head gesture* and *blink* and is on par on *saccade* and *3D gaze*. It is only worse w.r.t. SOTA on *Affective*, with a 6% decrease.

**Cross-dataset.** VAT, ChildPlay, and GFIE are challenging in-the-wild datasets (see Appendix B), making them well suited for generalization evaluation. For instance, on VAT and ChildPlay, we cannot report results for the 1DCNN models of [86], as MediaPipe [12] often fails under extreme head poses. In *Gesture*, OmniHead performs slightly below SOTA on KTH in STL, although it remains competitive with MTL for both macro $F_1^{F_{ma}}$ and $F_1^{E_{ma}}$. This is expected, as KTH's meeting scenarios are most similar to the CCDb training data. However, OmniHead generalizes substantially better to in-the-wild conditions, achieving improvements of 19% and 13% on VAT and ChildPlay in STL and MTL, respectively. On *Gaze*, they clearly improve performance on blink, while remaining on par with SOTA for saccade and 3D gaze. Finally, OmniHead$_{\text{MTL}}$ is within 3% of OmniHead$_{\text{STL}}$, indicating that the unified model preserves robustness across datasets.

| Ablations | | Gesture | | | | | | Gaze | | | | | | Affective | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Token Decoder | Pretraining | CCDb ↻ Head Gesture | | KTH ⇆ Head Gesture | | VAT ⇆ Head Gesture | | VAT ↻ Blink | ChildPlay ⇅ Blink | VAT ↻ Saccade | ChildPlay ⇅ Saccade | Gaze360 ↻ 3D Gaze | GFIE ⇆ 3D Gaze | sAff-Wild2 ↻ Expression | sAff-Wild2 ↻ Valence-Arousal | sAff-Wild2 ↻ Action Units |
| | | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^{E_{mi}}$ | $F_1^{E_{ma}}$ | $F_1^{E}$ | $F_1^{E}$ | $F_1^{E}$ | $F_1^{E}$ | Full | Full | $F_1^{F_{ma}}$ | CCC | $F_1^{F_{ma}}$ |
| STL | | | | | | | | | | | | | | | | |
| | | 0.71 | 0.57 | 0.49 | 0.35 | 0.46 | 0.20 | 0.58 | 0.34 | 0.56 | 0.50 | 11.69 | 20.27 | 0.23 | 0.39 | 0.43 |
| | ✓ | 0.73 | 0.65 | **0.61** | **0.52** | **0.61** | **0.44** | 0.62 | **0.46** | 0.68 | **0.61** | 12.03 | 21.37 | **0.34** | 0.45 | **0.45** |
| ✓ | | 0.70 | 0.59 | 0.53 | 0.39 | 0.55 | 0.20 | 0.60 | 0.35 | 0.57 | 0.46 | **11.52** | **18.96** | 0.31 | 0.36 | **0.45** |
| ✓ | ✓ | **0.74** | **0.67** | 0.54 | 0.44 | **0.61** | 0.43 | **0.67** | 0.44 | **0.70** | **0.61** | 11.78 | 20.65 | **0.34** | **0.46** | **0.45** |
| MTL | | | | | | | | | | | | | | | | |
| | | 0.72 | 0.63 | 0.30 | 0.20 | 0.48 | 0.19 | 0.61 | 0.34 | 0.63 | 0.50 | 12.18 | 26.02 | 0.27 | 0.23 | 0.40 |
| | ✓ | **0.73** | 0.63 | **0.60** | 0.36 | 0.58 | 0.28 | **0.65** | 0.46 | **0.68** | **0.60** | **12.08** | 21.85 | 0.28 | 0.37 | **0.46** |
| ✓ | | 0.71 | 0.62 | 0.34 | 0.22 | 0.51 | 0.18 | 0.60 | 0.35 | 0.61 | 0.49 | 12.19 | 22.17 | 0.21 | 0.23 | 0.40 |
| ✓ | ✓ | 0.72 | **0.65** | 0.52 | **0.50** | **0.60** | **0.34** | 0.62 | **0.48** | 0.67 | 0.59 | 12.50 | **21.44** | **0.30** | **0.44** | 0.44 |

**Table 5 Ablations.** Impact of the task token decoder and pretraining in both STL and MTL settings. $\text{OmniHead}_{\text{STL}}$ and $\text{OmniHead}_{\text{MTL}}$ correspond to the last rows. ↻ within and ⇆ cross dataset evaluation.

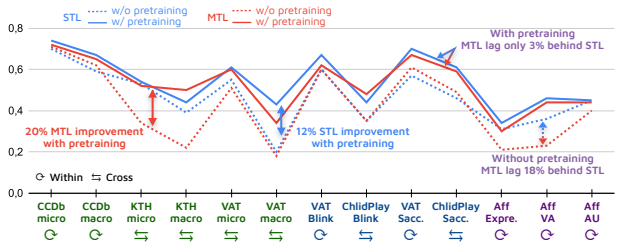



**Figure 3 Ablation visualizations** of Tab. 5 pretraining impact (rows 3, 4, 7, 8). Reported percentages indicate average relative improvements across all tasks (including 3D gaze).

**Pretraining.** In Tab. 4, our explicit semi-supervised representation learning approach substantially outperforms task-agnostic pretraining methods. We finetune MARLIN [13], VideoMAE [81], and OmniHead in both STL and MTL settings. The results are clear: OmniHead achieves an overall improvement of 23% over MARLIN and 9% over VideoMAE in STL, and 23% and 19%, respectively, in MTL. Interestingly, although MARLIN is pretrained on facial videos, the larger scale and diversity of the datasets used in VideoMAE appear to yield more effective representations for dynamic facial behavior recognition.

## 5.4. Ablation Study

Additional ablations are discussed in the appendix, such as temporal encoding (Sec. G.1), multi-task weight loss (Sec. G.2) and tasks relationships (Sec. G.3). Qualitative examples are presented in the supplement video.

**What is the impact of our semi-supervised pretraining?** As shown in Tab. 5 and Fig. 3, in STL, pretrained models achieve an average improvement of 12% across all tasks and datasets ($3^{th}$ vs $4^{th}$ row), leading to more robust representations less prone to overfitting: they better generalize (16% gain in cross-dataset vs 9% in within-dataset), and also benefit the recognition of underrepresented classes, *e.g.* in head gesture, the macro $F_1^{E_{ma}}$ improves by 26%, compared to 6% for $F_1^{E_{mi}}$, with similar trends observed for blink and saccade detection. Only the gaze angular error shows a slight degradation, likely because pretraining may bias the model to $\pm 90^o$ frontal faces compared to $\pm 180^o$ in Gaze360 and GFIE.

In MTL, jointly training diverse nonverbal behaviors across multiple in-the-wild datasets is substantially more challenging than single-task learning, particularly without pretraining. As shown in Fig. 3, this results in an average relative performance drop of 18% compared to STL models. Pretraining mitigates this gap, reducing it to only 3%, and thus proves essential for learning a robust unified representation that aligns all tasks. During pretraining, each sample is supervised across all tasks, whereas in the available datasets, each sample is typically annotated for only a single task.

Moreover, in MTL, pretraining the unified model yields an overall improvement of 20%, as shown in Fig. 3, with gains of 14% in the within-dataset and 29% in the cross-dataset settings. These results further demonstrate the model's ability to enhance generalization across challenging datasets.

**Does the task token decoder improve performance?** As a baseline, we remove the transformer decoder and use one MLP per task, taking each frame feature map average pooling as input. In Tab. 5, in STL without pretraining, the task token decoder yields a slight improvement of 4% overall ($1^{th}$ vs $3^{th}$ row). This improvement is reduced when the model is pretrained. In MTL with pretraining, the task token decoder shows a slight improvement of 3%, suggesting that a task token decoder brings benefit in multi-task setting such as task token specialization and inter-task sharing.

## 5.5. Limitations and Future Work

Besides head gestures, most behaviors are observable only when the face is visible. Although saccades and gaze direction can sometimes be inferred from head pose and motion, modeling face visibility may be necessary to suppress face-related predictions during face occlusion.

Some behaviors, such as affective and 3D gaze, can be inferred from an image. Restricting the training of spatiotemporal models to video only reduces dataset availability and diversity. Recent works address this by training spatiotemporal models with both video and image jointly [5, 27, 28].

# 6. Conclusion

This work introduced OmniHead, the first unified spatiotemporal framework for dynamic nonverbal facial behavior analysis. OmniHead jointly detects head gestures, blinks, saccades/fixations, 3D gaze, facial expressions, valence/arousal, and action units. It further addresses, for the first time, the detection of head gestures, blinks, and saccades/fixations in the wild, providing dense annotations on two challenging datasets. In addition, the proposed semi-supervised pretraining strategy substantially enhances generalization, particularly on underrepresented classes, narrowing the gap with single-task learning and surpassing state-of-the-art single-task models across multiple tasks.

# OmniHead: A Unified Model for Dynamic Nonverbal Facial Behaviors

## Supplementary Material

## A. What the Supplementary Material Provides

OmniHead produces predictions for seven tasks, making it difficult to visualize in a figure. We therefore include a **video of qualitative results** in the supplementary material that shows results on a wide variety of samples. The supplementary document also provides additional details on the **Datasets** B, **Annotations** C, **Baselines** D, **Pseudo-labels** E, **Losses/training** F and **Experiments** G.

## B. Dataset Details

**CelebV-HQ [96].** CelebV-HQ is a large-scale facial video dataset containing 68 hours of in-the-wild footage. It includes 35,666 clips from 15,653 identities, offering high diversity in both facial appearance and dynamics. Each clip consists of a fixed head crop at a high spatial resolution of $512 \times 512$ pixels. Although the dataset captures various head poses, most are near-frontal (typically within a $\pm 80^\circ$ range). Compared to VoxCeleb2 [21], CelebV-HQ is smaller in overall duration but provides a greater number of identities, wider head pose distribution, and higher resolution, making it a strong candidate for dynamic head representation learning.

**CCDb-HG [86].** CCDb-HG extends the original Cardiff Conversation Database (CCDb) [6] with dense and comprehensive gesture annotations. CCDb consists of 49 non-scripted dyadic conversations recorded from a frontal viewpoint, focusing on upper-body motion to study facial backchannel behaviors and gestures. CCDb-HG includes 4,731 annotated head gestures: 2,469 nods, 848 shakes, 523 tilts, 643 turns, and 238 up/down gestures. Training and testing are performed using a subject-independent split, where four subjects (25 videos) are held out for testing out of the total 115 videos.

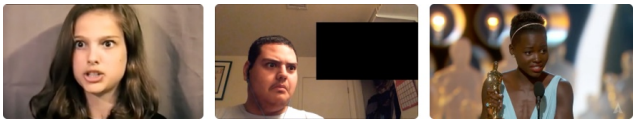
**KTH-Idiap [59].** The KTH-Idiap dataset is relatively small and used only for evaluation. It features groups of four individuals engaged in discussions around a table, leading to greater head motion and fewer frontal views compared to CCDb-HG. In total, KTH-Idiap contains nine videos (one per participant).

**Gaze360 [35].** Gaze360 is a large-scale video dataset for 3D gaze estimation, recorded in both indoor and outdoor environments under unconstrained conditions. It contains 3D gaze annotations for 238 subjects with wide variations in head pose and gaze direction. The dataset is recorded at 8 FPS. In all experiments, we follow the official training and testing splits from [35], using 126,928 training samples and 25,969 test samples (referred to as "All 360°" in [35]).

**GFIE [33].** GFIE is an indoor 3D gaze dataset containing 71,799 frames from 61 subjects (27 male, 34 female). It captures natural gaze behavior across a wide range of head poses. Using a calibrated laser setup, 3D gaze vectors from the eye to the target are precisely measured. Recordings were collected at 30 FPS while participants performed various indoor activities. We follow the data splits from [33], comprising 59,217 training, 6,281 validation, and 6,281 test samples.

**Video Attention Target (VAT) [20].** VAT is a video dataset constructed from high-resolution clips extracted from popular TV shows. It was originally designed for the gaze-following task and exhibits a broader head orientation distribution compared to typical facial analysis datasets. VAT contains 606 clips from 50 different shows, totaling 71,666 frames with 949 head tracks in the training set, 86 in validation, and 298 in testing. Although diverse, the scenes remain limited in scope due to their TV-based origin.

**ChildPlay [80].** ChildPlay is a recent video dataset built from YouTube videos and annotated for the gaze-following task, focusing on children's gaze behavior. Compared to VAT, ChildPlay features more challenging viewpoints, lower head resolution, and greater visual variability. It also provides valuable data on children, who are underrepresented in most existing datasets. From 95 source videos, 401 clips were extracted, totaling 120,549 frames, with 734 head tracks for training, 63 for validation, and 118 for testing.



**s-Aff-Wild2 [39].** s-Aff-Wild2 is a static subset of the original Aff-Wild2 [40] dataset, designed for the multi-task Affective Behavior Analysis in-the-Wild (ABAW) Challenge [39]. It provides 142,382 training images and 26,876 validation images; the test set is not publicly released. To prevent hyperparameter selection on the official validation set, we created an internal validation subset from the training data. Valence and arousal values range within $[-1, 1]$. The dataset includes 8 expression categories (Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Other) and 12 action units (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26). In this work, we use head crops rather than face crops to increase robustness to extreme head poses. Since Aff-Wild2 provides only cropped face images without corresponding bounding boxes in the raw videos, we first apply a head detector on the original videos. For clips with multiple detected heads, we extract face recognition embeddings for all detected heads as well as the provided face crop, and we select the head that has the maximum cosine similarity with the face crop.

## C. Annotation Details

Annotations were performed by paid bachelor's students trained on progressively harder samples until consistent performance. Videos were annotated in LabelBox at the frame level, and all annotations were reviewed by the authors. Quality was assessed via double annotation and inter-annotator agreement (Table S1). The gap between frame- and event-level metrics reflects slight temporal boundary variability.

## D. Baseline and Additional Methods

### D.1. Saccade Baseline

Existing approaches for saccade detection typically rely on threshold-based algorithms applied to gaze-direction trajectories. Following [79], we implement a velocity-based method that detects saccades from changes in the 3D gaze direction. The velocity threshold is selected by maximizing the event-level F1 score on the VAT validation set.

| | *Blink* | | | *Saccade* | | | *Head Gesture* | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | CK | $F_1^F$ | $F_1^E$ | CK | $F_1^F$ | $F_1^E$ | CK | $F_1^F$ | $F_1^E$ |
| VAT | 0.58 | 0.61 | 0.78 | 0.63 | 0.69 | 0.84 | 0.64 | 0.67 | 0.81 |

**Table S1 Inter-annotator agreement on VAT test set.** frame-based: Cohen Kappa (CK) and $F_1^F$, event-based: $F_1^E$

We evaluate two variants of the same gaze estimator [85], which can operate either on static images or on short video clips. As shown in Tab. S2, we apply the threshold-based baseline to both the image-based and video-based 3D gaze predictions. The threshold determined on VAT is then used unchanged for cross-dataset evaluation on ChildPlay. Results show that the video-based predictions are smoother and lead to substantially better saccade detection performance than the image-based predictions.

| | *Gaze Behavior - Saccade* | | | |
|---|---|---|---|---|
| | VAT ↻ | | ChildPlay ⇆ | |
| Saccade Baseline | $F_1^F$ | $F_1^E$ | $F_1^F$ | $F_1^E$ |
| w/ image gaze prediction | 0.59 | 0.68 | 0.43 | 0.57 |
| w/ video gaze prediction | 0.64 | 0.69 | 0.50 | 0.62 |

**Table S2 Saccade Baseline.** Impact of image vs video-based gaze prediction for the saccade threshold-based baseline method. ↻ within and ⇆ cross datasets evaluation

### D.2. Blink Baseline

Existing blink detection methods are also limited and often depend on facial landmarks or eye crops, as in MediaPipe. However, such approaches are not robust to extreme head poses, making them unsuitable for our setting. For example, in the Video Attention Target (VAT) dataset, MediaPipe frequently fails to detect faces. We develop a simple algorithm based on an observation. Gaze predictions differ between image-based and video-based models during blinking events. Specifically, video models produce smoother gaze trajectories, while image-based models tend to predict a downward jitter, as closed eyelids are visually similar to looking down (e.g. in Fig. 1 frame 1 and 2 are visually similar, but the person blinks in the first frame). Based on this discrepancy during blinking, we design a threshold-based algorithm that detects blinks from the angular difference between 3D gaze predictions obtained from video and image models. The threshold is optimized by maximizing the event-level F1 score on the validation set. This baseline performs reliably under near-frontal views but degrades as head orientation becomes more extreme.

### D.3. Head Gesture 1D-CNN Flame

The original method proposed in [86] relies on features extracted using MediaPipe [12]. As discussed above, MediaPipe often fails under challenging in-the-wild conditions. To address this limitation, we extract equivalent features using VGGHead [42], a model that predicts FLAME parameters and is robust to extreme head poses. From these parameters, we derive 3D landmarks and head pose features, which we use to retrain the gesture model following the procedure in [86]. We extract these new features on the CCDb-HG [86] dataset and retrain the model accordingly. As shown in Tab. 2 and Tab. 3, the updated model achieves slightly higher accuracy than the MediaPipe-based baseline (1D-CNN [86]), although feature reliability still decreases for extreme head poses due to temporal jitter.

## E. Pseudo-label Extraction Details

As described in the main paper, we pretrain OmniHead on CelebV-HQ using expert-model distillation and therefore require high-quality pseudo-labels.

**Gesture Behavior.** CelebV-HQ contains challenging head-pose variations and diverse illumination, requiring a robust gesture estimator. We use the 1D-CNN FLAME model introduced in Sec. D.3, which provides reliable head gesture predictions in in-the-wild conditions. We further filter low-confidence predictions using probability thresholding and temporal smoothing.

**Gaze Behavior.** Given the variability in CelebV-HQ, robust 3D gaze estimation is essential. We employ the ST-WSGE Gaze Transformer [85], trained on Gaze360 and GazeFollow, which shows strong in-the-wild performance. The model operates on single images or 8-frame clips. We found the video model with a stride of one yields the smoothest results and we use it to extract 3D gaze pseudo-labels. Using these gaze estimates, we then apply our saccade and blink baselines described in Secs. D.1 and D.2 to extract saccade and blink pseudo-labels. We apply temporal smoothing and enforce a minimum event duration of three frames for both behaviors.

**Affective Behavior.** For facial expression, valence/arousal, and action units, we use the expert models referenced in the main paper. For expression and action units, we do not apply hard thresholding and instead treat the model outputs as soft labels during supervision.

## F. Objective Functions and Training

### F.1. Classification

**Head Gesture, Facial Expression, Blink, and Saccade.** Head-gesture labels include *none* (background), *nod*, *shake*, *tilt*, *turn*, and *down/up*. Facial-expression labels follow Aff-Wild2 and include *Neutral*, *Anger*, *Disgust*, *Fear*, *Happiness*, *Sadness*, *Surprise*, and *Other*. Blink labels are *no_blink* and *blink*, and saccade labels are *fixation* and *saccade*. All four tasks are trained with a cross-entropy loss:

$$L_{\text{task}} = -\sum_{i=1}^{N_{\text{class}}} c_i \, p_i \log \hat{p}_i, \tag{S1}$$

where $p_i = 1$ if the sample belongs to class $i$ and 0 otherwise, $\hat{p}_i$ denotes the predicted probability after the softmax, and $c_i$ is the class weight.

For head gestures, we apply a label-smoothing factor of $0.15$. For expression recognition on Aff-Wild2, we use class weights $[0.47, 2.34, 3.34, 3.74, 0.62, 1.41, 2.08, 0.46]$ derived from the inverse class distribution to address class imbalance. For blink and saccade on VAT, we use class weights $[1, 10]$ and apply a label-smoothing factor of $0.1$.

**Action Unit.** For facial action unit, it includes AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, AU26. We used the binary cross-entropy loss function:

$$\mathcal{L}_{au} = -\sum_{i=1}^{12}[(1-p_i)\log(1-\hat{p}_i) + p_i \log \hat{p}_k \tag{S2}$$

where $p_i = 1$ for the i-th action unit if it exists and 0 otherwise. $p_i = 1$ is the predicted i-th action unit output after the sigmoid function.

### F.2. Regression

**Geometric losses.** During pretraining, the model predicts the FLAME parameters. Using a FLAME layer [46], these parameters are converted into 3D landmarks, 3D vertices, and a head-pose rotation matrix. Following VGG-Head [42], we regress these outputs using three losses. (1) Reprojection loss: measures the discrepancy between the projected 3D vertices and the pseudo 2D keypoint coordinates.

$$\mathcal{L}_{land} = \frac{1}{N_{coord}} \sum_{i=1}^{N_{coord}} ||v_i - \hat{v}_i||_1 \tag{S3}$$

where $N_{coord}$ is the number of keypoints. We only use the facial keypoints and not all the head keypoints. 2) the vertices loss: we calculate the L2 Loss over the normalized and unrotated 3D Head Vertices. The global rotation predictions are set to zero to evaluate the discrepancy between our predictions and the pseudo-vertices in 3D.

$$\mathcal{L}_{vert} = \frac{1}{N_{coord}} \sum_{i=1}^{N_{coord}} ||v_i|_{R=0} - \hat{v}_i||_2 \tag{S4}$$

3) Rotation loss: the geodesic distance loss is used which is specific to measure matrix discrepancies:

$$\mathcal{L}_{pose} = \cos^{-1}\left(\frac{tr(R_p R_{gt}^T) - 1}{2}\right) \tag{S5}$$

**Valence/Arousal.** Following [72], we use the consistency correlation coefficient that measures the agreement between two variables and ranges from -1 to 1, with higher values indicating better agreement, defined as:

$$CCC(X, \hat{X}) = \frac{2COV(X, \hat{X})}{\delta_X^2 + \delta_{\hat{X}}^2 + (\mu_X - \mu_{\hat{X}})^2} \tag{S6}$$

where $\delta_X$ is the variances and $\mu_X$ is the mean and $COV$ the covariance. Therefore, we defined the loss as

$$\mathcal{L}_{va} = 1 - CCC(va, \hat{va}) + 1 - CCC(ar, \hat{ar}) \tag{S7}$$

**3D Gaze.** For gaze estimation we follow [85] minimizing the angular error defined as:

$$\mathcal{L}_{gaze} = \frac{180}{\pi} \arccos\left(\frac{g^T \hat{g}}{||g||_2 ||\hat{g}||_2}\right) \tag{S8}$$

### F.3. Training details

Models are optimized using AdamW with a learning rate of $1e{-4}$ with cosine decay, and weight decay of $1e{-3}$. Pretraining is performed for 20 epochs on four RTX 3090 GPUs using distributed data parallelism. OmniHead is then trained for up to 25 epochs on a single RTX 3090 GPU with early stopping based on validation performance. During finetuning, the encoder learning rate is reduced to $5e{-5}$. Since each sample has 16 frames, the batch size per GPU is set to 12.

We apply standard data augmentations during pretraining and finetuning, including horizontal flips, color jitter, and Gaussian blur. We also introduce temporal jitter by shifting the input window by a few frames and adjusting the corresponding labels. During pretraining, multitask loss weights are set to 1 for regression tasks and 10 for classification tasks. For training OmniHead, task-specific loss weights are set to 5 for gaze and blink, 2 for head gestures, 1 for saccades, and 0.1 for affective tasks.

## G. Additional Experiments

### G.1. Spatio-Temporal Encoder

We previously argued in Sec. 4.1.1 that the encoder must capture subtle temporal variations, including head motion and rapid gaze shifts. Temporal information can be incorporated at different stages of the architecture. Late temporal fusion applies a temporal model on top of a spatial encoder (e.g., DINO+GRU). Early temporal encoding, in contrast, integrates temporal cues directly from the input, as in video Transformers such as VideoSwin, which perform spatiotemporal self-attention on input patches. Other approaches adapt image Transformers by inserting temporal-processing layers throughout the network. ST-Adapter [62], for example, adds a temporal convolution before the spatial self-attention in each block, enabling DINO+ST-Adapter to encode temporal information earlier than DINO+GRU.

| | *Gesture* | | | | *Gaze* | | | |
|---|---|---|---|---|---|---|---|---|
| | **CCDb** ↻ *Head Gesture* | | **KTH** ⇆ *Head Gesture* | | **VAT** ↻ *Saccade* | **ChildPlay** ⇆ *Saccade* | **Gaze360** ↻ *Gaze* | **GFIE** ⇆ *Gaze* |
| Encoder | $F_1^{Emi}$ | $F_1^{Ema}$ | $F_1^{Emi}$ | $F_1^{Ema}$ | $F_1^{E}$ | $F_1^{E}$ | Full | Full |
| STL | | | | | | | | |
| $\text{Dino}_{\text{freeze}}$+GRU | 0.41 | 0.32 | 0.20 | 0.11 | 0.50 | 0.37 | 19.89 | 24.62 |
| Dino+GRU | 0.37 | 0.32 | 0.18 | 0.05 | 0.56 | 0.48 | 11.66 | 18.88 |
| $\text{Dino}_{\text{freeze}}$+ST-adapter [62] | 0.64 | 0.49 | 0.42 | 0.15 | 0.51 | 0.36 | 13.17 | 20.44 |
| Dino+ST-adapter [62] | 0.61 | 0.51 | 0.39 | 0.25 | 0.62 | 0.52 | 12.33 | 19.13 |
| VideoSwin | 0.71 | 0.57 | 0.49 | 0.35 | 0.56 | 0.50 | 11.69 | 20.27 |

**Table S3 Temporal Encoder Comparison.** Using OmniHead with a simple MLP decoder, we investigate the impact of different temporal encoders from late (DINO+GRU) to early (Dino+ST-adapter, VideoSwin) temporal encoding. ↻ within and ⇆ cross datasets evaluation

We evaluate these design choices on tasks that rely on temporal dynamics, head gestures and saccades, and include 3D gaze estimation as an image-level task. Our hypothesis is that subtle motion patterns require early temporal encoding (DINO+ST-Adapter or VideoSwin), while late temporal fusion (DINO+GRU) may suffice when temporal cues are weaker.

**Head gestures.** Results in Tab. S3 show that head-gesture recognition strongly benefits from early temporal encoding. DINO+ST-Adapter and VideoSwin substantially outperform DINO+GRU across datasets. Fine-tuned DINO+ST-Adapter achieves a 24% absolute improvement in event micro-$F_1$ on CCDDb compared with DINO+GRU, with similar gains on KTH and under other metrics. These results indicate that preserving short-term motion cues—akin to optical flow—is essential, and early temporal encoders capture these cues more effectively.

**Saccades.** For saccade detection, the trend is weaker but consistent. DINO+ST-Adapter outperforms DINO+GRU by 6% on VAT, whereas VideoSwin yields no measurable gain. This suggests that early temporal encoding can help with fine-grained saccade dynamics, but late temporal fusion already captures most of the saccadic events.

**3D gaze.** As expected, temporal encoding plays a limited role in 3D gaze estimation. Performance differences between early (VideoSwin) and late (DINO+GRU) fusion strategies are negligible on Gaze360. The benefit of using temporal information in this task stems primarily from temporal smoothing rather than from modeling subtle motion patterns, and both types of encoders support this equally well.

| | *Gesture* | | | | *Gaze* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **CCDb** ⟲ *Head Gesture* | | **KTH** ⇆ *Head Gesture* | | **VAT** ⟲ *Blink* | **ChildPlay** ⇆ *Blink* | **VAT** ⟲ *Saccade* | **ChildPlay**$^{cross}$ *Saccade* | **Gaze360** ⟲ *Gaze* | **GFIE** ⇆ *Gaze* |
| OmniHead | $F_1^{Emi}$ | $F_1^{Ema}$ | $F_1^{Emi}$ | $F_1^{Ema}$ | $F_1^{E}$ | $F_1^{E}$ | $F_1^{E}$ | $F_1^{E}$ | Full | Full |
| STL | | | | | | | | | | |
| Pretrained w/ adaptive weight [36] | 0.70 | 0.59 | 0.53 | 0.34 | 0.61 | **0.47** | 0.68 | 0.59 | **11.70** | **20.59** |
| Pretrained w/ selected weight | **0.74** | **0.67** | **0.54** | **0.44** | **0.67** | 0.44 | **0.70** | **0.61** | 11.78 | 20.65 |
| No pretraing MTL (HG,Bli.,Sac.,Gaze) | | | | | | | | | | |
| w/ adaptive weights [36] | 0.71 | **0.60** | 0.34 | 0.16 | 0.59 | 0.36 | 0.60 | 0.52 | 12.63 | 24.84 |
| w/ selected weights | **0.72** | 0.59 | **0.38** | **0.19** | **0.62** | **0.40** | **0.62** | **0.53** | **12.10** | **23.68** |

**Table S4 Multi-task adaptive weight loss** In STL, during pretraining, we tried w/ and w/o automatic adaptive weighted method [36]. In MTL, without pretraining, we tried w/ and w/o automatic adaptive weighted method to learn *head gesture*, *blink*, *saccade*, and *gaze* tasks jointly. ⟲ within and ⇆ cross datasets evaluation

## G.2. Multi-task Weighted Loss

Multi-task learning seeks to improve efficiency and accuracy by optimizing several objectives within a shared representation. Joint optimization, however, is challenging because tasks may converge at different rates or require distinct feature abstractions. The standard formulation uses a weighted sum of task-specific losses, but selecting appropriate weights is difficult in practice. Several studies address this issue through adaptive optimization strategies. Kendall *et al*. [36] derive a weighting scheme based on homoscedastic task uncertainty, enabling the model to learn suitable loss weights for both regression and classification objectives.

We apply this adaptive weighting strategy in two settings: (i) during pretraining, and (ii) during multi-task learning (MTL) of OmniHead$_{\text{MTL}}$.

**Pretraining.** We evaluate downstream single-task finetuning (OmniHead$_{\text{STL}}$) after pretraining with and without adaptive weighting. In the non-adaptive setting, weights are manually tuned. As shown in Tab. S4, rows 1–2, adaptive weighting yields lower downstream performance. The degradation is particularly pronounced for head gestures: adaptive weighting increases the emphasis on auxiliary geometric objectives during pretraining, which appears to hinder the learning of motion-sensitive behaviors in subsequent finetuning.

**MTL.** We further examine adaptive weighting when training OmniHead$_{\text{MTL}}$ jointly on head gestures, blinks, saccades, and gaze. Results in Tab. S4, rows 3–4, show that adaptive weighting again produces slightly lower performance overall. This suggests that the proposed framework is relatively robust to the choice of task-loss weights and does not benefit from uncertainty-based weighting in this context.

## G.3. Tasks Relationships

One hypothesis is that multi-task learning (MTL) can be beneficial since they are naturally interconnected. The results do not support this hypothesis yet. Nevertheless, task

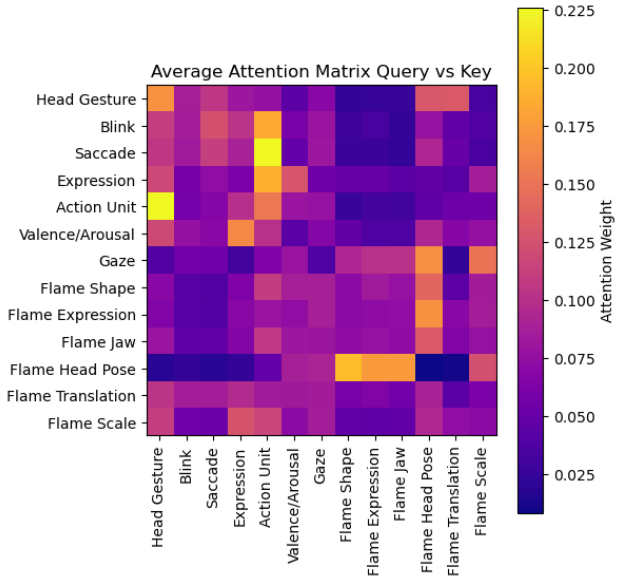



**Figure S1 Averaged Attention Weight after pretraining.** Visualization of the attention weight from the task-token self-attention layer in the decoder's last block for the middle frame, averaged over the attention heads and over 5k randomly selected samples from CelebV-HQ test set.

relationships can be explored from our learned MTL representation in the task self-attention, the main mechanism for inter-task communication. Fig. S1 visualizes the averaged attention matrix and reveals interesting patterns that give insight into task relationships:(i) AUs inform blink, saccade, and expression, consistent with their role as facial movement primitives (FACS [22]); (ii) head pose is the strongest support for gaze prediction, (iii) head gestures modulate multiple tasks; (iv) affective tasks (expression, AU, valence/arousal) are closely linked.

## References


[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023.

[2] Andra Adams, Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. Decoupling facial expressions and head motions in complex emotions. In *2015 International conference on affective computing and intelligent interaction (ACII)*, pages 274–280. IEEE, 2015.

[3] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[4] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 25–32, 2014.

[5] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[6] Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282, 2013.

[7] Valentina Bachurina and Marie Arsalidou. Multiple levels of mental attentional demand modulate peak saccade velocity and blink rate. *Heliyon*, 8(1), 2022.

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

[9] Tadas Baltrušaitis, Amir Zadeh, Yao Chong, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018, IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2016.

[10] Paris Mavromoustakos Blom, Sander Bakkes, Chek Tan, Shimon Whiteson, Diederik Roijers, Roberto Valenti, and Theo Gevers. Towards personalised gaming via facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 30–36, 2014.

[11] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. In *European Conference on Computer Vision*, pages 107–125. Springer, 2022.

[12] H. Nash C. McClanahan E. Uboweja M. Hays F. Zhang C.-L. Chang M. Yong J. Lee W.-T. Chang W. Hua M. Georg C. Lugaresi, J. Tang and M. Grundmann. Mediapipe: A framework for perceiving and processing reality. In *In Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2019.

[13] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1493–1504, 2023.

[14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[15] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[16] C Chen, Y. Yu, and J.-M. Odobez. Head nod detection from a full 3d model. In *Int. Conf. on Computer Vision Workshop, Santiago, Chile,*, 2015.

[17] Chu-Song Chen, Hsuan-Tien Lin, et al. 360-degree gaze estimation in the wild using multiple zoom scales. In *British Machine Vision Conference (BMVC)*, 2021.

[18] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.

[19] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022.

[20] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5396–5406, 2020.

[21] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[22] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[23] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, and Klaus Scherer. Relative importance of face, body, and speech in judgments of personality and affect. *Journal of personality and social psychology*, 38(2):270, 1980.

[24] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.

[25] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, pages 334–352, 2018.

[26] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, Seyedmohammad Mavadati, and Dean P Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[27] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16102–16112, 2022.

[28] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10406–10417, 2023.

[29] Yiran Guan, Zhuoguang Chen, Wenzheng Zeng, Zhiguo Cao, and Yang Xiao. End-to-end video gaze estimation via capturing head-face-eye spatial-temporal interaction context. *IEEE Signal Processing Letters*, 30:1687–1691, 2023.

[30] Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. *Advances in Neural Information Processing Systems*, 37:15646–15673, 2024.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[32] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1439–1449, 2021.

[33] Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8907–8916, 2023.

[34] Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*, pages 126–142. Springer, 2022.

[35] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019.

[36] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[37] Daeha Kim and Byung Cheol Song. Optimal transport-based identity matching for identity-invariant facial expression recognition. *Advances in neural information processing systems*, 35:18749–18762, 2022.

[38] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

[39] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2022.

[40] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[41] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9980–9989, 2021.

[42] Orest Kupyn, Eugene Khvedchenia, and Christian Rupprecht. Vggheads: 3d multi head alignment with a large-scale synthetic dataset. *arXiv preprint arXiv:2407.18245*, 2024.

[43] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013.

[44] R John Leigh and David S Zee. *The neurology of eye movements*. Oxford university press, 2015.

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

[46] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[47] Xiaodong Li, Wenchao Du, and Hongyu Yang. Affective behavior analysis using task-adaptive and au-assisted graph. In *European Conference on Computer Vision*, pages 393–403. Springer, 2024.

[48] Jing Liang, Yu-Qing Zou, Si-Yi Liang, Yu-Wei Wu, and Wen-Jing Yan. Emotional gaze: The effects of gaze direction on the perception of facial emotions. *Frontiers in psychology*, 12:684357, 2021.

[49] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),*, 43(3):1092–1099, 2021.

[50] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[51] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[52] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1239–1246, 2022.

[53] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.

[54] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. Facial features for affective state detection in learning environments. In *Proceedings of the annual meeting of the cognitive science society*, 2007.

[55] Skanda Muralidhar, Rémy Siegfried, Jean-Marc Odobez, and Daniel Gatica-Perez. Facing employers and customers: What do gaze and expressions tell about soft skills? In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, MUM 2018, Cairo, Egypt, November 25-28, 2018*, pages 121–126, 2018.

[56] Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11369–11382, 2025.

[57] Dang-Khanh Nguyen, Sudarshan Pant, Ngoc-Huynh Ho, Guee-Sang Lee, Soo-Hyung Kim, and Hyung-Jeong Yang. Affective behavior analysis using action unit relation graph and multi-task cross attention. In *European Conference on Computer Vision*, pages 132–142. Springer, 2022.

[58] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2192–2201, 2022.

[59] Catharine Oertel, Kenneth A Funes Mora, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. Who will get the grant? a multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 workshop on understanding and modeling multiparty, multimodal interactions*, pages 27–32, 2014.

[60] Kazuhiro Otsuka and Masahiro Tsumori. Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, 8:217169–217195, 2020.

[61] Patrizia Paggio, Manex Agirrezabal, Bart Jongejan, and Costanza Navarretta. Automatic detection and classification of head movements in face-to-face conversations. In *Proceedings of LREC2020 Workshop" People in language, vision and the mind"(ONION2020)*, pages 15–21, 2020.

[62] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning. In *Advances in Neural Information Processing Systems*, pages 26462–26477. Curran Associates, Inc., 2022.

[63] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2223–2234, 2023.

[64] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *European Conference on Computer Vision*, pages 240–260. Springer, 2024.

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

[66] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017.

[67] R. Reiter-Palmon, T. Sinha, J. Gevers, J.-M. Odobez, and G. Volpe. Theories and models of teams and group. *Journal of Small Group Research*, 45(5):544–567, 2017.

[68] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 41–50, 2021.

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

[70] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th international symposium on intelligent systems and informatics (SISY)*, pages 119–124. IEEE, 2021.

[71] Andrey V Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2359–2366, 2022.

[72] Andrey V Savchenko. Hsemotion team at the 7th abaw challenge: multi-task learning and compound facial expression recognition. *arXiv preprint arXiv:2407.13184*, 2024.

[73] S. Sheikhi and J.M. Odobez. Combining dynamic head pose and gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters*, 66:81–90, 2015.

[74] R. Siegfried and J.-M. Odobez. Robust unsupervised gaze calibration using conversation and manipulation attention priors. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(1):1–27, 2022.

[75] Rémy Siegfried, Yu Yu, and Jean-Marc Odobez. A deep learning approach for robust head pose independent eye movements recognition from videos. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 31:1–31:5, New York, NY, USA, 2019. ACM.

[76] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 147–152. IEEE, 2013.

[77] Valeriya Strizhkova, Laura M Ferrari, Hadi Kachmar, Antitza Dantcheva, and François Brémond. Video representation learning for conversational facial expression recognition guided by multiple view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4693–4702, 2024.

[78] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Maedfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023.

[79] Qi Sun, Anjul Patney, Li-Yi Wei, Omer Shapira, Jingwan Lu, Paul Asente, Suwen Zhu, Morgan McGuire, David Luebke, and Arie Kaufman. Towards virtual reality infinite walking: dynamic saccadic redirection. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[80] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946, 2023.

[81] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[82] Jessica L Tracy and David Matsumoto. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences*, 105(33):11655–11660, 2008.

[83] Alexandria K Vail, Tadas Baltrušaitis, Luciana Pennant, Elizabeth Liebson, Justin Baker, and Louis-Philippe Morency. Visual attention in schizophrenia: Eye contact and gaze aversion during clinical interactions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 490–497. IEEE, 2017.

[84] Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing gaze estimation with weak-supervision from synthetic views. In *ECCV*, 2024.

[85] Pierre Vuillecard and Jean-Marc Odobez. Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13508–13518, 2025.

[86] Pierre Vuillecard, Arya Farkhondeh, Michael Villamizar, and Jean-Marc Odobez. Ccdb-hg: Novel annotations and gaze-aware representations for head gesture recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE, 2024.

[87] Lingfeng Wang, Haocheng Li, and Chunyin Liu. Hybrid cnn-transformer model for facial affect recognition in the abaw4 challenge. *arXiv preprint arXiv:2207.10201*, 2022.

[88] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[89] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3601–3610, 2021.

[90] Chao Yan, Weiguo Pan, Cheng Xu, Songyin Dai, and Xuewei Li. Gaze estimation via strip pooling and multi-criss-cross attention networks. *Applied Sciences*, 13(10): 5901, 2023.

[91] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[92] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20291–20300, 2022.

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

[94] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–60, 2017.

[95] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022.

[96] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022.

[97] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020.