

# 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**.

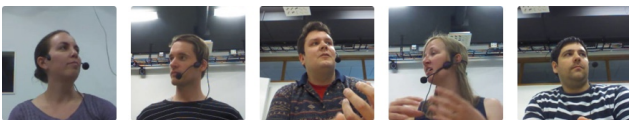
### B. Dataset Details



**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

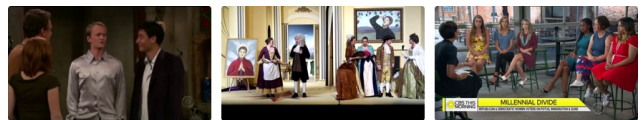
and used only for evaluation. It features groups of four individuals engaged in discussions around a table, leading to greater head motion and fewer frontal views compared to CCDb-HG. In total, KTH-Idiap contains nine videos (one per participant).



**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]).



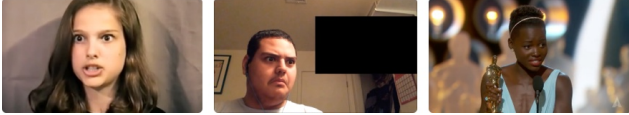
**GFIE [33]**. GFIE is an indoor 3D gaze dataset containing 71,799 frames from 61 subjects (27 male, 34 female). It captures natural gaze behavior across a wide range of head poses. Using a calibrated laser setup, 3D gaze vectors from the eye to the target are precisely measured. Recordings were collected 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

	<i>Blink</i>			<i>Saccade</i>			<i>Head Gesture</i>		
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:  $F_1^E$

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.

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.

	<i>Gaze Behavior - Saccade</i>			
	VAT $\odot$		ChildPlay $\rightleftharpoons$	
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.  $\odot$  within and  $\rightleftharpoons$  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*, *Hap-*

*piness*, *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:

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

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:

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

where  $p_i = 1$  for the  $i$ -th action unit if it exists and 0 otherwise.  $\hat{p}_i = 1$  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.

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

where  $N_{\text{coord}}$  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.

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

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

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



**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:

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

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

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

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

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

### F.3. Training details





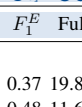
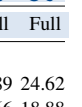
Models are optimized using AdamW with a learning rate of  $1e-4$  with cosine decay, and weight decay of  $1e-3$ . Pretraining is performed for 20 epochs on four RTX 3090 GPUs using distributed data parallelism. OmniHead is then trained for up to 25 epochs on a single RTX 3090 GPU with early stopping 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

	Gesture				Gaze			
	 				   			
Encoder	$F_1^{E_{mi}}$	$F_1^{E_{ma}}$	$F_1^{E_{mi}}$	$F_1^{E_{ma}}$	$F_1^E$	$F_1^E$	Full	Full
STL								
Dino <sub>freeze</sub> +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
Dino <sub>freeze</sub> +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.  $\odot$  within and  $\leftrightarrow$  cross datasets evaluation

self-attention in each block, enabling DINO+ST-Adapter to encode temporal information earlier than DINO+GRU.

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 $\odot$ Head Gesture		KTH $\Rightarrow$ Head Gesture		VAT $\odot$ Blink		ChildPlay $\Rightarrow$ Blink		
	$F_1^{E_{mi}}$	$F_1^{E_{ma}}$	$F_1^{E_{mi}}$	$F_1^{E_{ma}}$	$F_1^E$	$F_1^E$	$F_1^E$	$F_1^E$	
OmniHead									Full Full
STL									
Pretrained w/ adaptive weight [36]	0.70	0.59	0.53	0.34	0.61	<b>0.47</b>	0.68	0.59	<b>11.70 20.59</b>
Pretrained w/ selected weight	<b>0.74</b>	<b>0.67</b>	<b>0.54</b>	<b>0.44</b>	<b>0.67</b>	0.44	<b>0.70</b>	<b>0.61</b>	11.78 20.65
No pretraing MTL (HG,Bli.,Sac.,Gaze)									
w/ adaptive weights [36]	0.71	<b>0.60</b>	0.34	0.16	0.59	0.36	0.60	0.52	12.63 24.84
w/ selected weights	<b>0.72</b>	0.59	<b>0.38</b>	<b>0.19</b>	<b>0.62</b>	<b>0.40</b>	<b>0.62</b>	<b>0.53</b>	<b>12.10 23.68</b>

**Table S4 Multi-task adaptive weight loss** In STL, during pre-training, 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.  $\odot$  within and  $\Rightarrow$  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.

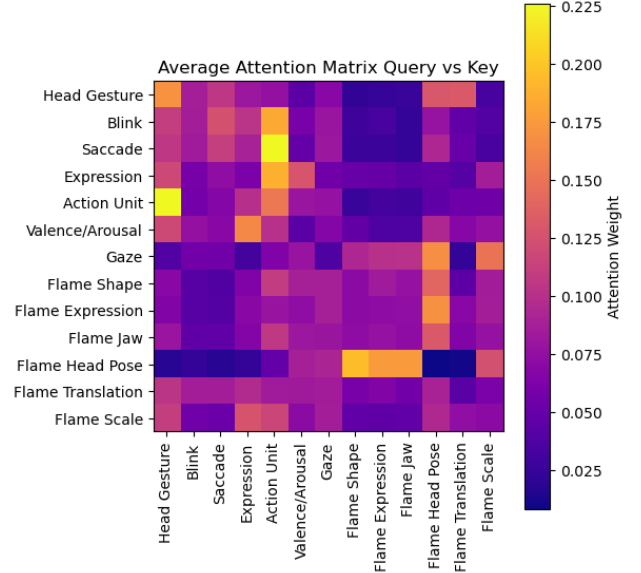
We apply this adaptive weighting strategy in two settings: (i) during pretraining, and (ii) during multi-task learning (MTL) of OmniHead<sub>MTL</sub>.

**Pretraining.** We evaluate downstream single-task finetuning (OmniHead<sub>STL</sub>) 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<sub>MTL</sub> 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. 6
- [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. 1
- [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. 1
- [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. 1
- [5] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 1
- [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. 6, 1
- [7] Valentina Bachurina and Marie Arsalidou. Multiple levels of mental attentional demand modulate peak saccade velocity and blink rate. *Heliyon*, 8(1), 2022. 4
- [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. 2, 3
- [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. 2, 3
- [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. 1
- [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. 2, 3
- [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. 5, 7, 3
- [13] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezaatofighi, 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. 2, 3, 4, 5, 7, 1
- [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. 3
- [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. 3
- [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*. 3
- [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. 2, 7
- [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. 2
- [19] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022. 2, 4, 6
- [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. 2, 4, 6, 1
- [21] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 1
- [22] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 4, 5
- [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. 1
- [24] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013. 1
- [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. 2
- [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. 1

- [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. 4, 6, 1
- [28] Rohit Girdhar, Alaeldin 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. 4, 1
- [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. 2, 7
- [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. 2
- [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. 3
- [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. 6
- [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. 6, 1
- [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. 1
- [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. 2, 6, 7, 1
- [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. 5
- [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. 2, 5
- [38] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986. 1
- [39] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision*, pages 157–172. Springer, 2022. 3, 5, 6, 7, 2
- [40] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 3, 5, 6, 2
- [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. 2, 7
- [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. 5, 6, 3
- [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. 2
- [44] R John Leigh and David S Zee. *The neurology of eye movements*. Oxford university press, 2015. 4
- [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. 2
- [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. 4, 6, 3
- [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. 3, 6, 7
- [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. 1, 2
- [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. 2
- [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. 4
- [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. 4, 5
- [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. 2, 4, 5



- [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. 4
- [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. 1
- [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. 1
- [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. 2, 3, 4
- [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. 3, 5, 7
- [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. 2
- [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. 6, 1
- [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. 1, 2, 7
- [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. 2, 7
- [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. 4
- [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. 2, 3
- [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. 2, 3, 4
- [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. 3
- [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. 2, 3
- [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. 1
- [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. 1
- [69] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 4
- [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. 2, 4, 5
- [71] Andrey V Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using efficient-nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2359–2366, 2022. 7
- [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. 3, 6, 7, 4
- [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. 1
- [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. 2
- [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. 3
- [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. [2](#)
- [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. [3](#)
- [78] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023. [3](#)
- [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. [2](#)
- [80] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Child-play: A new benchmark for understanding children’s gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20935–20946, 2023. [2](#), [4](#), [6](#)
- [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. [3](#), [4](#), [7](#), [1](#)
- [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. [1](#)
- [83] Alexandria K Vail, Tadas Baltrušaitis, Luciana Penant, 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. [1](#), [2](#)
- [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. [2](#), [6](#)
- [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. [2](#), [4](#), [5](#), [6](#), [7](#), [3](#)
- [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. [2](#), [3](#), [5](#), [6](#), [7](#), [1](#)
- [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. [6](#), [7](#)
- [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. [2](#)
- [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. [2](#)
- [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. [2](#)
- [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. [2](#)
- [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. [2](#)
- [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. [5](#)
- [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. [2](#)
- [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. [2](#), [3](#), [4](#), [5](#)
- [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. [3](#), [6](#), [1](#)
- [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. [2](#)