CVPR #6241

AVF-MAE++: Scaling Affective Video Facial Masked Autoencoders via Efficient Audio-Visual Self-Supervised Learning

Anonymous CVPR submission

Paper ID 6241

Abstract

Affective Video Facial Analysis (AVFA) is important for advancing emotion-aware AI, yet the persistent data scarcity in AVFA presents challenges. Recently, the self-supervised learning (SSL) technique of Masked Autoencoders (MAE) has gained significant attention, particularly in its audio-visual adaptation. Insights from general domains suggest that scaling is vital for unlocking impressive improvements, though its effects on AVFA remain largely unexplored. Additionally, capturing both intra- and inter-modal correlations through robust representations is a crucial challenge in this field. To tackle these gaps, we introduce AVF-MAE++, a series audio-visual MAE designed to explore the impact of scaling on AVFA with a focus on advanced correlation modeling. Our method incorporates a novel audio-visual dual masking strategy and an improved modality encoder with a holistic view to better support scalable pre-training. Furthermore, we propose the Iteratively Audio-Visual Correlations Learning Module to improve correlations capture within the SSL framework, bridging the limitations of prior methods. To support smooth adaptation and mitigate overfitting, we also introduce a progressive semantics injection strategy, which structures training in three stages. Extensive experiments across 17 datasets, spanning three key AVFA tasks, demonstrate the superior performance of AVF-MAE++, establishing new state-of-the-art results. Ablation studies provide further insights into the critical design choices driving these gains. The code will be released soon.

1. Introduction

Affective Video Facial Analysis (AVFA) aims to detect and interpret human affective states from facial videos, which has great application values in fields such as HCI [50] and dialogue systems [56]. Since audio-visual cues (*e.g.*, facial expressions and prosody) predominantly contribute to 93% of emotional perceptions [46, 72], audio-visual AVFA has made rapid progress over the past decades. With the fast development of deep learning and numerous datasets, super-

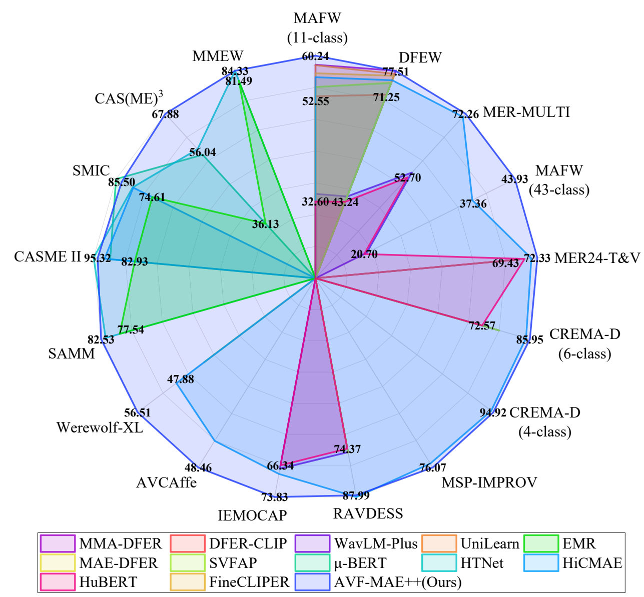

Figure 1. Performance comparisons of AVF-MAE++ and state-of-the-art AVFA methods on 17 datasets across CEA, DEA, and MER tasks. Notably, we report the averaged results over dimensions on both Werewolf-XL [79] and AVCAffe [55] datasets.

vised deep models have been the mainstream paradigm for AVFA [29, 80, 81]. Although supervised learning has made great strides, it fatally requires large-scale labeled data. Furthermore, it is extremely expensive and time-consuming to annotate high-quality emotions [73].

A natural intuition is to utilize abundant unlabeled video data to compensate for the AVFA data scarcity. As a result, self-supervised learning (SSL) methods for AVFA has drawn massive attention, particularly Masked Autoencoders (*i.e.*, MAE) [28]. Specifically, MAE aims to reconstruct the raw data from masked facial videos, leading to the emergence of various visual and audio-visual AVFA MAE methods [58, 60, 61]. Meanwhile, following the promising findings in image and language domains [2, 28], VideoMAE V2 [67] has shown that scaling model capacity and data size are essential for exhibiting remarkable performance gains. However, very few work has explored the scaling properties of MAE pre-training for AVFA, which is more severe in audio-visual field. While [60, 61] provides models with varying capacities, their largest size generally reaches the ten-million scale, lagging behind those in general domains. More importantly, a key aspect for scalable audio-visual MAE pre-training is the effective capture of intra- and inter-modal correlations through robust representations since the prevalent audio-visual co-expressions of emotions. Nonetheless, existing AVFA methods under SSL manner still have limitations in capturing correlated cues [73, 89].

To fill these gaps, we aim to explore the scaling properties of audio-visual MAE for AVFA, with a focused emphasis on capturing intra- and inter-modal correlations, pushing the performance limits across diverse downstream datasets. Building upon HiCMAE [60], we scale the audio-visual MAE and further conduct million-level data scaling for the pre-training stage to harness their full potential. In addition, we design related components to explicitly enhance the capture of audio-visual correspondences, addressing the dilemma faced by existing AVFA methods. However, we still need to carefully tackle several issues as below:

(1) Computational costs and memory consumptions remain the primary bottlenecks in scaling audio-visual MAE. Although [60] adopts the asymmetric encoder-decoder design from [62], it still struggles to fully support the pre-training for large-scale models. Inspired by the dual masking strategy for the asymmetric architecture from [67], we adaptively present the audio-visual dual masking strategy, leading to a more efficient audio-visual self-supervised proxy task. Meanwhile, the vanilla global space-time attention mechanism incurs quadratic scaling costs, and large redundancy (*e.g.*, facial symmetry) exists in 3D facial video data, rendering the expenses suboptimal. We thus flexibly introduce a local-global interaction attention paradigm for modality encoders, while elevating the holistic view to compensate for its weakness in global information flow. (2) A huge number of unlabeled data is still required to facilitate scalable pre-training. Unlike affective analysis in images, existing AVFA datasets are typically smaller in scale. To tackle this, a simple solution is to mix the unlabeled video data from multiple sources. Following [58, 60], we mainly mix datasets towards speaker recognition and successfully build a large-scale pre-training dataset with around 1.36M clips. (3) Previous methods commonly utilize self-attention and cross-attention components to build correlations modeling layers, leading to inadequate cross-modal interactions and a lack of hierarchical aggregative integrations across multi-semantic scales. Besides, they often neglect the role of multi-modal features in learning comprehensive representations. To this end, we propose the IAV-CL Module (**I**teratively **A**udio-**V**isual **C**orrelations **L**earning Module), which effectively promotes the capture of audio-visual correlations. (4) A key challenge for SSL methods is smoothly adapting the pre-trained models to downstream datasets. Directly performing fine-tuning on small-scale downstream datasets often leads to severe overfitting, hindering the full potential of pre-trained models. Therefore, we propose the progressive semantics injection (PSI) strategy that leverages supervised hybrid datasets from diverse sources to act as a bridge between pre-training and downstream fine-tuning.

Based on the above analysis, we propose a series audio-visual MAE termed AVF-MAE++. By leveraging the dual scaling in both model capacity and data size, along with the introduced IAV-CL Module, we further present the PSI strategy to construct a three-stage progressive training pipeline. The overall pipeline consists of large-scale audio-visual masked pre-training, post-pretraining on supervised hybrid datasets, and targeted fine-tuning on downstream datasets. To verify the effectiveness of our method, we conduct extensive pre-training and evaluate model performance across three key downstream tasks involving 17 datasets. As illustrated in Fig. 1, our AVF-MAE++ outperforms various state-of-the-art supervised or self-supervised approaches. Remarkably, AVF-MAE++ is the first method to surpass 60% WAR on MAFW (11-class) [42] dataset.

In addition to advancing AVFA, our work also contributes to research in MLLMs [13], talking face generation [68], and deepfake detection [69]. Our main contributions are three-fold: (1) We introduce AVF-MAE++ to explore the scaling properties of audio-visual MAE for AVFA, incorporating efficient dual scaling and a progressive adaptation strategy. As pioneers, we strive to lay a solid foundation for future research. (2) Departing from previous methods, we adaptively introduce a local-global interaction attention paradigm enhanced with a more holistic perspective. We further propose the IAV-CL Module to explicitly improve the capture of intra- and inter-modal correlations. (3) Extensive experiments across 17 datasets, spanning three AVFA tasks, verify the effectiveness of AVF-MAE++. We also justify the design choices of our method by ablations.

2. Related Work

Audio-visual AVFA. Most studies on audio-visual AVFA fall into the supervised learning paradigm [23, 73], primarily focusing on two important aspects: uni-modal feature extraction and audio-visual information fusion. With the progress of deep learning, various video and audio feature extractors have been developed [43, 66, 82]. Recently, the success of self-supervised learning in general domains has spurred the emergence of large pre-trained models, achieving significant results in emotion analysis [30, 58, 61]. Regarding audio-visual information fusion [45, 60], model-level fusion is the most widely adopted strategy, mainly building upon the self-attention and cross-attention components [59, 64, 81]. Despite promising results, most of the above audio-visual AVFA methods are under the supervised learning manner, which are severely constrained by the scarcity of labeled emotion data and domain shifts.

Masked Audio-Visual Modeling. Masked data modeling learns representations by reconstructing the masked portions of input. Previous works [25, 33, 71] have extended this learning approach to the audio-visual domain, demonstrating impressive results across a variety of downstream tasks. Among them, MAE-style methods have attracted significant interests due to their efficient data learning capabilities [22, 24, 47]. However, the representations learned by these methods are typically unsuitable for AVFA, as they are not specifically trained on facial video data. Recently, VQ-MAE-AV [54] introduces a vector-quantized MAE tailored for audio-visual AVFA, and Sun et al. [60] present HiCMAE with a three-pronged learning strategy. Despite these advancements, the scaling properties of MAE-style methods have not been thoroughly explored for audio-visual AVFA, leaving substantial gaps. In addition, there is still room for improvement in the correlations capture of above methods. In this paper, we introduce AVF-MAE++, aiming to bridge these gaps and promote the progress of AVFA.

Masked Autoencoders Scaling. Building upon the foundational success of MAE, researchers have widely explored its scaling properties across various fields. SimMIM [75] studies the data scaling capability of masked image modeling. VideoMAE [62] and MAE-ST [21] have trained the huge video transformers with millions of parameters, while VideoMAE V2 [67] scales the VideMAE [62] in both model capacity and data size. Han et al. [26] propose the *Efficient* MAE with a novel loss and a new masking strategy. Singh et al. [57] present an additional pre-pretraining stage to improve model initialization. In AVFA, some works have initially explored the scaling properties of MAE [58, 60, 61], but they primarily focus on limited model scaling, with minimal exploration of data scaling. To this end, we scale audio-visual MAE in terms of both model capacity and data size with the currently largest AVFA pre-training dataset.

3. Methodology

In this section, we begin by revisiting the foundational work HiCMAE [60] in Sec. 3.1. We then introduce the audio-visual dual masking strategy (Sec. 3.2), improved modality encoder (Sec. 3.3), and the IAV-CL Mdoule (Sec. 3.4), as shown in Fig. 2. Finally, we elaborate on the details of dual scaling and the progressive adaptation strategy (Sec. 3.5).

3.1. HiCMAE Revisited

HiCMAE [60] follows the asymmetric encoder-decoder architecture of [28] and proposes a three-pronged hierarchical strategy. Next, we briefly revisit its implementation details.

Data Embedding. A cube embedding layer and a patch embedding layer are first utilized to divide $\mathbf{X}_v \in \mathbb{R}^{T_v \times H \times W \times 3}$ and $\mathbf{X}_a \in \mathbb{R}^{T_a \times F}$, leading to token lists: $\mathbf{X}_v^{'} = \Phi_{emb}^{v}(\mathbf{X_v})$ and $\mathbf{X}_a^{'} = \Phi_{emb}^{a}(\mathbf{X_a})$, where $\mathbf{X}_v^{'} = \{\mathbf{X}_v^{'i}\}_{i=1}^{N_v}$ and $\mathbf{X}_a^{'} = \{\mathbf{X}_a^{'j}\}_{j=1}^{N_a}$ are the token sequences, $(\mathbf{X}_v^{'i}, \mathbf{X}_a^{'j}) \in \mathbb{R}^{1\times C}$ are the tokens output by the embedding layers and then added with positional embeddings. Here, $N_v = \frac{T_v}{2} \times \frac{H}{16} \times \frac{W}{16}$ and $N_a = \frac{T_a}{16} \times \frac{F}{16}$ refer to the lengths of video and audio token sequences, while C denotes the feature channels.

Token Masking. HiCMAE deploys the tube masking and random masking for the video and audio branches, using high masking ratios ($\rho_v = 90\%$ and $\rho_a = 80\%$). Next, only the visible tokens $\mathbf{X}_v^{''}$ and $\mathbf{X}_a^{''}$ will run into the encoder, where $\mathbf{X}_v^{''} = \{\mathbf{X}_v^{'i}\}_{i\in(1-\mathbb{M}(\rho_v))}$, $\mathbf{X}_a^{''} = \{\mathbf{X}_a^{'j}\}_{j\in(1-\mathbb{M}(\rho_a))}$, and their token lengths are $N_v^{\prime} = 0.1N_v$ and $N_a^{\prime} = 0.2N_a$. $\mathbb{M}(\rho_v)$ and $\mathbb{M}(\rho_a)$ here are the audio-visual masking maps.

Encoder. The encoder of HiCMAE simply operates on the visible tokens $\mathbf{X}_v^{''}$ and $\mathbf{X}_a^{''}$ with two modality-specific encoders and a cross-modal fusion encoder: $\mathbf{E}_{a\to v}$, $\mathbf{E}_{v\to a}$ $= \Phi_{enc}^{a\leftrightarrow v}(\Phi_{enc}^{v}(\mathbf{X}_v^{''}), \Phi_{enc}^{a}(\mathbf{X}_a^{''}))$, where the modality encoders are vanilla ViT [17], and the fusion encoder is mainly implemented using multi-head cross-attention components.

Decoder. The video and audio decoders, including hierarchical skip connections, respectively take the *combined* tokens as inputs and reconstruct data with narrower and shallower ViT: $\hat{\mathbf{X}}_m = \Phi_{dec}^{m}(\mathbf{E}_m^c)$, where the *combined* tokens $\mathbf{E}_m^c$ is the concatenated sequence of encoded tokens $\mathbf{E}_{\bar{m}\to m}$ and the learnable masked tokens $[\mathtt{MASK}]_m$ (with position embeddings), the token length $N_m^d = N_m$, and $m \in \{a, v\}$.

Pre-training Loss. The pre-training object is to minimize the combination of modality-specific *Mean Square Error* (MSE) Losses and the introduced *HCMCL* Loss [60], *i.e.*,

$$\mathcal{L} = (\mathcal{L}_{\text{MSE}}^{a} + \mathcal{L}_{\text{MSE}}^{v}) + \lambda \cdot \sum_{k=1}^{N_c} \mathcal{L}_{\text{InfoNCE}}(\mathbf{e}_a^k, \mathbf{e}_v^k), \tag{1}$$

where λ is the weight factor, N_c is the number of selected enocder layers in hierarchical skip connections, and $\mathbf{e}_m^k$ is a batch of sample-level features to adopt HCMCL Loss.

Downstream Fine-tuning. After pre-training, the overall encoder incorporating hierarchical feature fusion will be deployed to targetedly fine-tune on the downstream tasks.

3.2. Audio-Visual Dual Masking Strategy

As analyzed in Sec. 3.1, the decoders of HiCMAE need to process the overll tokens, leading to large redundancy. Recently, VideoMAE V2 [67] introduces the dual masking strategy, where the decoder takes inputs from the visible tokens under the encoder mask $\mathbb{M}_e = \mathcal{M}_e(\rho^e)$ and part of the remaining tokens visible under the decoder mask $\mathbb{M}_d = \mathcal{M}_d(\rho^d)$, leading to more efficient video pre-training.

Inspired by this insight, we present the audio-visual dual masking strategy, including encoder masking $\mathcal{M}_e$ and decoder masking $\mathcal{M}_d$ for both audio and video branches, as illustrated in Fig. 2 (a). Specifically, the encoder masking $\mathcal{M}_e^m$ keeps consistent with HiCMAE [60]. For $\mathcal{M}_d^v$, we follow [67] to adaptively adopt the running cell masking [52]

CVPR #6241

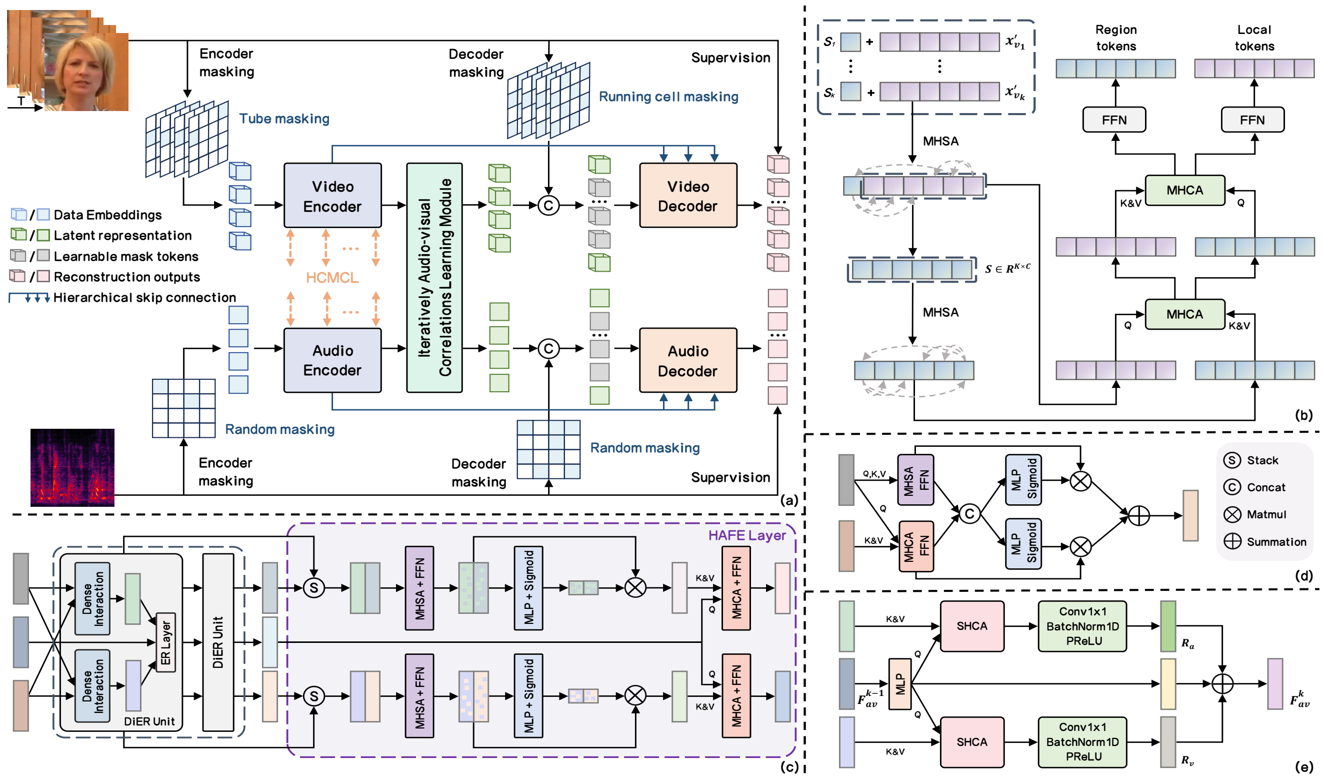

Figure 2. The overall illustrations of AVF-MAE++. (a) The pre-training pipeline with our new audio-visual dual masking strategy. (b) The improved modality encoder. (c) IAV-CL Module. (d) & (e) The *dense interaction* and *evolutionary refinement* layers of one DiER Unit.

to boost information complement in this partial reconstruction. Regarding $\mathcal{M}_d^a$, we also deploy the random masking since the prior knowledge in [32] indicates that audio MAE learns easily by predicting nearby contexts. Following [67], ρ^d of $\mathcal{M}_d$ for both audio and video branches are 50%. With this introduced dual masking strategy, the new *combined* token sequence for modality decoder can be formulated as:

$$\mathbf{E}_m^c = \mathbf{E}_{\bar{m}\to m} \cup \{\mathbf{M}_i^m\}_{i\in \mathbb{M}_d^m}, \tag{2}$$

where $\mathbf{E}_{\bar{m}\to m}$ denotes the latent features from encoder, $\mathbf{M}_i^m$ is the learnable masking token with related positional embeddings, and $m \in \{a, v\}$. With this updated sequence $\mathbf{E}_m^c$, decoder only regards the visible tokens as the reconstruction targets. The final MSE Loss can be given as:

$$\mathcal{L}_{\text{MSE}}^m = \frac{1}{(1-\rho_m^d)\cdot N_m} \sum_{i\in \mathbb{M}_d^m \cap \mathbb{M}_e^m} |\mathbf{X}_m^i - \hat{\mathbf{X}}_m^i|^2, \tag{3}$$

where $\mathbf{X}_m$ and $\hat{\mathbf{X}}_m$ denotes the original input and the reconstructed output of audio-visual modalities, respectively.

3.3. Improved Modality Encoder

The large redundancy in facial videos, coupled with the heavy computations of global space-time self-attention in vanilla ViT [17], impedes efficient large-scale pre-training. Motivated by this, we adaptively adopt and improve the LGI-Former [58] for the uni-modal encoder since its effectiveness in reducing computational costs. For simplicity, we describe only one encoder layer during fine-tuning, which mainly differs from the pre-training stage in the number of visible tokens per region. The original LGI-Former [58] is proposed for video, which can be decomposed into three stages: (I) local intra-region self-attention, (II) global inter-region self-attention, and (III) local-global interaction.

In stage I, the 3D tokens $\mathbf{X}_v^{'} \in \mathbb{R}^{\frac{T_v}{2}\times\frac{H}{16}\times\frac{W}{16}\times C}$ is first divided into K non-overlapping local spatio-temporal regions of equal size $Z_v = t\times h\times w$, leading to $\mathbf{X}_{v_i}^{'} \in \mathbb{R}^{Z_v\times C}$, and $\mathbf{X}_{v_i}^{'}$ is then added with a learnable region token $\mathbf{S}_i \in \mathbb{R}^{1\times C}$ ($i \in \{1, 2, ..., K\}$, $K = \frac{N_v}{Z_v}$). The self-attention then operates on their concatenation to promote local-aware features learning and aggregate information into the region token $\mathbf{S}_i$:

$$\hat{\mathbf{X}}_{v_i}^{'} = \text{MHSA}(\text{LN}(\text{C}(\mathbf{S}_i, \mathbf{X}_{v_i}^{'}))) + \text{C}(\mathbf{S}_i, \mathbf{X}_{v_i}^{'}), \tag{4}$$

where $\hat{\mathbf{X}}_{v_i}^{'} \in \mathbb{R}^{(Z_v+1)\times C}$, MHSA$(\cdot)$ is the vanilla multi-head self-attention, LN$(\cdot)$ and C$(\cdot)$ denote layer normalization and concatenation operation. In stage II, all the region tokens $\{\mathbf{S}_i\}_{i=1}^K$ are first aggregated, self-attention is then employed to exchange inter-region information between different regions with negligible costs, *i.e.*,

$$\mathbf{S} = \text{MHSA}(\text{LN}(\text{C}(\mathbf{S}_1, ..., \mathbf{S}_K))) + \text{C}(\mathbf{S}_1, ..., \mathbf{S}_K), \tag{5}$$

where $\mathbf{S} \in \mathbb{R}^{K\times C}$ is the aggregated region tokens. So far, the region token $\mathbf{S}_i$ has been consolidated by discriminative information from other regions, holding a global perspective of the overall input tokens. As a result, in stage III, multi-head cross-attention between $\mathbf{X}^{'}_{v_i}$ and $\mathbf{S}$ is explicitly exploited to enable the original local tokens to access the global-aware selective information, *i.e.*,

$$\mathbf{X}^{'}_{v_i} = \text{MHCA}(\text{LN}(\mathbf{X}^{'}_{v_i}), \text{LN}(\mathbf{S})) + \mathbf{X}^{'}_{v_i}, \tag{6}$$

where MHCA$(\cdot)$ refers to the vanilla multi-head cross-attention. Additionally, since the region tokens are important to the global evolutionary information flow across multiple encoder layers, which should emphasize the more holistic master for the local-aware intra-region information, we thus introduce the stage IV (global-local interaction):

$$\mathbf{S} = \text{MHCA}(\text{LN}(\mathbf{S}), \text{LN}(\mathbf{X}^{'}_{v_i})) + \mathbf{S}. \tag{7}$$

Subsequently, both local and region tokens run through the feed-forward networks (FFNs) to perform further refinements. When applying this encoder to the audio branch, the main difference is the patch embedding layer outputs 2D tokens $\mathbf{X}^{'}_{a} \in \mathbb{R}^{\frac{T_a}{16} \times \frac{F}{16} \times C}$, leading to different region shape. In stage I, we split $\mathbf{X}^{'}_{a}$ into K non-overlapping local regions of equal size $Z_a = h_a \times w_a$, resulting in $\mathbf{X}^{'}_{a_i} \in \mathbb{R}^{Z_a \times C}$ ($i \in \{1, 2, ..., K\}$, $K = \frac{N_a}{Z_a}$). After region division, the remaining process keeps consistent with the video branch. Finally, we take the region tokens $(\mathbf{S}_v, \mathbf{S}_a) \in \mathbb{R}^{K \times C}$ as the outputs from one encoder layer, and the overall modality encoders both consists of N_l sequentially stacked layers.

3.4. Iteratively Audio-Visual Correlations Learning

As illustrated in Fig. 2 (c), we present the IAV-CL Module, which incorporates the Dense Interactions and Evolutionary Refinement (DiER) Units, as well as the Hierarchical Aggregations and Feedback Enhancement (HAFE) Layer, aiming to iteratively capture the complementary correlations.

During fine-tuning, we first stack $\{\mathbf{S}^n_v, \mathbf{S}^n_a\}^{N_l}_{n=1}$, leading to the uni-modal features $(\mathbf{F}_v, \mathbf{F}_a) \in \mathbb{R}^{K \times N_l \times C}$. We then utilize the learnable layer weights to dynamically unify features across different encoder layers followed by concatenation to output the original multi-modal feature $\mathbf{F}^0_{av}$, *i.e.*,

$$\mathbf{F}^0_{av} = \text{C}(\sum_{l=1}^{N_l} \alpha^a_l \mathbf{F}^l_a, \sum_{l=1}^{N_l} \alpha^v_l \mathbf{F}^l_v), \tag{8}$$

where $\mathbf{F}^0_{av} \in \mathbb{R}^{K \times 2C}$, $\sum_{l=1}^{N_l} \alpha^m_l = 1$. We then simply use poolings to reshape $\mathbf{F}_v$ and $\mathbf{F}_a$ as $(\mathbf{F}^1_v, \mathbf{F}^1_a) \in \mathbb{R}^{K \times C}$. The DiER Unit is proposed to perform dense audio-visual interactions and evolutionarily refine the multi-modal feature in the simultaneous manner, which is detailed as follows:

Dense Audio-Visual Interactions. Considering that sequentially connecting MHSA and MHCA blocks [12, 60] supports dense audio-visual interactions insufficiently, we adopt the parallel arrangement, as illustrated in Fig. 2 (d). Specifically, we first concatenate the parallel outputs of the complete attention blocks along the channel dimension, then compute the channel-wise attention scores to perform refinement via a linear layer and a sigmoid function. Next, we use summation to output the densely interacted features:

$$\mathbf{F}^2_m = \sigma\left(\mathbf{W}_s \mathbf{F}^{sc}_m + \mathbf{b}_s\right) \mathbf{F}^s_m + \sigma\left(\mathbf{W}_c \mathbf{F}^{sc}_m + \mathbf{b}_c\right) \mathbf{F}^c_m, \tag{9}$$

$$\mathbf{F}^s_m = \text{MHSA}\big(\text{LN}(\mathbf{F}^1_m)\big) + \mathbf{F}^1_m, \tag{10}$$

$$\mathbf{F}^c_m = \text{MHCA}\big(\text{LN}(\mathbf{F}^1_m),\ \text{LN}(\mathbf{F}^1_{\bar{m}})\big) + \mathbf{F}^1_m, \tag{11}$$

where $\mathbf{F}^{sc}_m = \text{C}\left(\mathbf{F}^s_m, \mathbf{F}^c_m\right)$, $\sigma(\cdot)$ is sigmoid function, $\mathbf{F}^2_m \in \mathbb{R}^{K \times C}$, $\mathbf{W}_*$ and $\mathbf{b}_*$ ($* \in \{s, c\}$) are learnable parameters.

Evolutionary Refinement (ER) Layer iteratively refines the multi-modal feature, leading to the following feedback enhancements of correlations capture. We first simply employ the linear layer to transform $\mathbf{F}^0_{av}$ into $\mathbf{F}^1_{av} \in \mathbb{R}^{K \times C}$, then attend $\mathbf{F}^1_{av}$ to audio-visual features using the single-head cross-attention (SHCA) block, which dynamically aggregates uni-modal useful information into $\mathbf{F}^1_{av}$, as illustrated in Fig. 2 (e). Inspired by [31], the convolutional block incorporating one 1×1 convolution followed by the Batch Normalization and PReLU sub-layers is then introduced to generate the residual features $\mathbf{R}_a$ and $\mathbf{R}_v$, which discriminatively learn invariant audio-visual representations, *i.e.*,

$$\mathbf{R}_m = \text{Conv}(\text{Att}(\mathbf{F}^1_{av}, \mathbf{F}^2_m, \mathbf{F}^2_m)), \tag{12}$$

where Conv$(\cdot)$ and Att$(\cdot)$ refer to the convolutional and SHCA blocks. Next, we sum multi-modal feature with $\mathbf{R}_m$ to produce features with highly correlated information:

$$\mathbf{F}^k_{av} = \text{LN}(\mathbf{F}^{k-1}_{av} + \mathbf{R}^{k-1}_a + \mathbf{R}^{k-1}_v), \tag{13}$$

where $k \in \{2, ..., N_c\}$ is the unit index. The parameters of each ER Layer are shared to facilitate evolutionary refinements. Finally, the outputs $\{\mathbf{F}^2_{a_i}, \mathbf{F}^2_{v_i}\}^{N_c}_{i=1}$ of all the units are preserved as features at multi-semantic scales, while $\mathbf{F}^{N_c}_{av}$ will be utilized for the following feedback enhancement.

We then present the HAFE Layer to hierarchically aggregate preserved features and promote correlated relationships modeling in reverse. Since features across units have distinct semantic scales, simply using poolings integrates the hierarchical representations inadequately. We thus first stack $\{\mathbf{F}^2_{m_i}\}^{N_c}_{i=1}$ along N_c to merge the scale-aware features, then deploy the unit-level MHSA followed by FFNs to provide aggregatively contextual integrations, which further considers the intra-modal correspondences, *i.e.*,

$$\gamma_m = \text{MHSA}(\text{LN}(\mathbf{F}^s_m) + \mathbf{F}^s_m, \tag{14}$$

where $\mathbf{F}^s_m = \text{Stack}(\mathbf{F}^2_{m_1}, ..., \mathbf{F}^2_{m_{N_c}})$. To select the most useful representations, we first apply the linear projection and sigmoid function to dynamically assign weights across different granularities. The weighted summation is then conducted to output the compatibly integrated features, *i.e.*,

$$\mathbf{F}^3_m = \sum_{l=1}^{N_c} \sigma\left(\mathbf{W}_{sf} \cdot \gamma^l_m + \mathbf{b}_{sf}\right) \cdot \gamma^l_m. \tag{15}$$

Stage	Task	Dataset	#Emos	Num	AC
Pre-training	–	Unlabeled Hybrid	–	1,360,531	Mix
Supervised Post-pre-training	–	CEA Labeled Hybrid	13	31,218	Mix
	–	MER Labeled Hybrid	3	1,007	Lab
Targeted Fine-tuning	CEA	MAFW [42]	11	9,172	Wild
			43	8,996	Wild
	CEA	DFEW [34]	7	11,697	Wild
	CEA	MER-MULTI [38]	6	3,784	Wild
	CEA	MER24-T&V [39]	6	5030	Wild
	CEA	IEMOCAP [4]	4	5,531	Lab
	CEA	CREMA-D [6]	6	7,442	Lab
			4	4,896	Lab
	CEA	RAVDESS [44]	8	1,440	Lab
	CEA	MSP-IMPROV [5]	4	7,798	Lab
	DEA	Werewolf-XL [79]	3	14,632	Lab
	DEA	AVCAffe [55]	2	58,112	Wild
	MER	SAMM [16]	3	133	Lab
	MER	CASME II [76]	3	145	Lab
	MER	SMIC [37]	3	164	Lab
	MER	CAS(ME)^3 [36]	3	943	Lab
	MER	MMEW [3]	3	300	Lab

Table 1. The statistics of data utilized for three training stages. AC: Acquisition Condition. Mix: Wild & Lab Environments.

Afterwards, we deploy MHCA followed by FFNs to facilitate the complementarily correlated information learning with $\mathbf{F}_{av}^{N_c}$ under feedback manner, which can be given as:

$$\mathbf{F}_m^4 = \mathrm{MHCA}(\mathrm{LN}(\mathbf{F}_m^3), \mathrm{LN}(\mathbf{F}_{av}^{N_c})) + \mathbf{F}_m^3. \quad (16)$$

Finally, we utilize poolings along the token dimension to reshape features as $\mathbf{F}_m^4 \in \mathbb{R}^C$, followed by concatenation and a specific linear layer to output the final results $\mathbf{F}_f$ of the overall tuned model. For the downstream classification and regression tasks, we respectively use the cross-entropy and mean square error losses. During pre-training, the main difference is the visible token number of input features.

3.5. Dual Scaling and Progressive Training

Model Scaling. The model capacity is the foremost force in improving performance. Following the scaling behaviors of [60, 62], we scale the capacity of AVF-MAE++ by constructing uni-modal encoders of varying dimensions, attention heads, and depths, leading to three versions (*i.e.*, Base, Large, and Huge), which are detailed in the supplementary material. The stacked number of our IAV-CL Module remains constant. Besides, we adhere to [60, 67] by using lightweight vanilla ViT [17] as decoder, while keeping the decoder capacity consistent across different model versions.
Data Scaling. We construct an unlabeled hybrid cross-linguistic facial video dataset to better support audio-visual MAE pre-training, originating from CN-Celeb series [20], MER2024 [39], VoxCeleb2-dev [15], AV-Speech [18], and CelebV-HQ [88], as illustrated in Tab. 1. After collection, we filter and crop videos using the pre-processing pipeline from [7] to reduce redundancy, resulting in a hybrid pre-training dataset with 1.36M clips. *To our knowledge, this is the largest dataset utilized for AVFA self-supervised pre-training*. More details are shown in supplementary material.
Progressive Adaptation Training. Compared to [62, 67], the non-overlapping data distributions between the pre-training and fine-tuning stages in AVFA, along with the limited fine-tuning data, lead to the adaptation and overfitting challenges, restricting the full potential of pre-trained models. To tackle this, inspired by [2, 67], we propose the progressive semantics injection (PSI) strategy, which incorporates supervised semantic signals from multiple sources to help pre-trained models gradually adapt to the downstream tasks, leading to a three-stage training pipeline. Concretely, we first conduct self-supervised pre-training on the unlabeled hybrid dataset. We then perform supervised post-pre-training on the labeled hybrid datasets to inject downstream semantics into pre-trained models. As displayed in Tab. 1, the labeled hybrid datasets are built by merging datasets for different downstream tasks and aligning their label semantics. Finally, we fine-tune models on targeted datasets to transfer the general semantics to task-specific knowledge.

4. Experiments

4.1. Downstream AVFA Tasks

To demonstrate the effectiveness and generalizability of the AVF-MAE++, we conduct extensive experiments on multiple datasets for three key AVFA tasks, as shown in Tab 1.
Categorical Emotion Analysis (CEA). CEA is the most common AVFA task, aiming to classify each sample into a predefined category. Following [60], we conduct detailed analysis on this task to explore the scaling properties of audio-visual MAE. We employ UAR, WAR, and WA-F1 as the metrics to evaluate performance across ten datasets.
Dimensional Emotion Analysis (DEA). DEA continuously represents the affective states, leading to more fine-grained emotional annotations. Following [60, 61], we utilize AVCAffe [55] and WereWolf-XL [79] to verify the superiority of AVF-MAE++. The evaluation metrics for [55] and [79] are WA-F1 and PCC, respectively. Besides, we have not built the labeled hybrid dataset for DEA since the disalignments of continuous emotional annotations.
Micro-Expression Recognition (MER). This task recognizes brief and subtle facial expressions that reveal hidden emotional states. In this paper, we deploy the UF1 metric to evaluate performance on five representive MER datasets.

4.2. Main Results

We transfer the pre-trained representations of AVF-MAE++ on 17 targeted datasets across three downstream AVFA tasks, as shown in Tab. 2. More comparisons and the implementation details are provided in supplementary material.
CEA. We compare with state-of-the-art CEA methods on ten datasets. We draw the following observations: (1) The SSL methods exhibit better performance compared to supervised methods due to their powerful and efficient capabilities in learning effective AVFA representations. (2) Audio-visual SSL methods generally surpass uni-modal SSL ones by leveraging the complementary correlations of

(a) MAFW (11-class)

Method	SSL	Mod.	#PS	UAR	WAR
HuBERT [30]	✓	A	95	25.00	32.60
WavLM-Plus [9]	✓	A	95	26.33	34.07
DFER-CLIP [85]	✓	V	153	39.89	52.55
SVFAP [61]	✓	V	78	41.19	54.28
MAE-DFER [58]	✓	V	85	41.62	54.31
UniLearn [11]	✓	V	101	43.72	58.44
T-MEP [81]	×	A+V	61	37.17	51.15
MMA-DFER [14]	✓	A+V	–	44.25	58.45
HiCMAE [60]	✓	A+V	81	42.65	56.17
AVF-MAE++ (B)	✓	A+V	169	43.10	57.50
AVF-MAE++ (L)	✓	A+V	303	45.36	59.13
AVF-MAE++ (H)	✓	A+V	521	**46.05**	**60.24**
FineCLIPER [8]	✓	T+V	20	45.01	56.91

(b) DFEW

Method	SSL	Mod.	#PS	UAR	WAR
WavLM-Plus [9]	✓	A	95	37.78	44.64
S2D [10]	✓	V	9	65.45	74.81
MAE-DFER [58]	✓	V	50	63.41	74.43
UniLearn [11]	✓	V	101	66.80	76.68
AMH [77]	×	A+V	–	54.48	66.51
HiCMAE [60]	✓	A+V	81	63.76	75.01
AVF-MAE++ (B)	✓	A+V	169	63.74	75.42
AVF-MAE++ (L)	✓	A+V	303	65.14	76.24
AVF-MAE++ (H)	✓	A+V	521	**66.88**	**77.45**
FineCLIPER [8]	✓	T+V	20	65.98	76.21

(c) MER-MULTI

Method	SSL	Mod.	#PS	UAR	WA-F1
HuBERT-CH [78]	✓	A	95	–	61.16
ResNet-FER [27]	×	V	26	–	57.44
MANet [86]	×	V	51	–	56.19
[27] + [78]	✓	A+V	121	–	69.11
[86] + [78]	✓	A+V	146	–	70.32
HiCMAE [60]	✓	A+V	81	64.15	71.33
AVF-MAE++ (B)	✓	A+V	169	64.87	69.56
AVF-MAE++ (L)	✓	A+V	303	66.34	70.79
AVF-MAE++ (H)	✓	A+V	521	**68.20**	**72.26**

(d) MAFW (43-class)

Method	SSL	Mod.	#PS	UAR	WAR
HuBERT [30]	✓	A	95	5.36	20.70
WavLM-Plus [9]	✓	A	95	5.51	21.09
Former-DFER [84]	×	V	18	10.21	32.07
T-MEP [81]	×	V	5	9.50	31.54
T-ESFL [42]	×	A+V	–	9.93	34.67
T-MEP [81]	×	A+V	61	13.22	36.58
HiCMAE [60]	✓	A+V	81	13.29	37.36
AVF-MAE++ (B)	✓	A+V	169	15.42	43.41
AVF-MAE++ (L)	✓	A+V	303	15.59	**43.93**
AVF-MAE++ (H)	✓	A+V	521	**17.25**	43.83

(e) MER24-T&V

Method	SSL	Mod.	#PS	WAR	WA-F1
Whisper [53]	×	A	1550	63.27	63.23
DINOv2 [49]	✓	V	–	59.57	58.44
VideoMAE [62]	✓	V	86	64.93	64.50
HiCMAE [60]	✓	A+V	81	70.95	70.18
AVF-MAE++ (B)	✓	A+V	169	72.11	71.24
AVF-MAE++ (L)	✓	A+V	303	**72.33**	71.64
AVF-MAE++ (H)	✓	A+V	521	72.28	**71.75**

(f) CREMA-D (6-class)

Method	SSL	Mod.	#PS	UAR	WAR
HuBERT [30]	✓	A	95	72.72	72.57
WavLM-Plus [9]	✓	A	95	73.34	73.39
SVFAP [61]	✓	V	78	77.31	77.37
MAE-DFER [58]	✓	V	85	77.33	77.38
VQ-MAE-AV [54]	✓	A+V	30	–	80.40
HiCMAE [60]	✓	A+V	81	84.91	84.89
AVF-MAE++ (B)	✓	A+V	169	85.10	85.09
AVF-MAE++ (L)	✓	A+V	303	85.69	85.60
AVF-MAE++ (H)	✓	A+V	521	**86.02**	**85.95**

(g) CREMA-D (4-class)

Method	SSL	Mod.	#PS	UAR	WAR
AW-HuBERT [63]	✓	A+V	103	93.65	93.65
HiCMAE [60]	✓	A+V	81	94.00	94.13
AVF-MAE++	✓	A+V	521	**94.82**	**94.92**

(h) MSP-IMPROV

Method	SSL	Mod.	#PS	UAR	WAR
FAV-HuBERT [63]	✓	A+V	103	61.05	68.35
HiCMAE [60]	✓	A+V	81	65.78	74.95
TAPT-HuBERT [63]	✓	A+V	103	63.95	70.46
AW-HuBERT [63]	✓	A+V	103	65.72	71.80
AVF-MAE++*	✓	A+V	521	**70.05**	**76.07**

(i) RAVDESS

Method	SSL	Mod.	#PS	UAR	WAR
HuBERT [30]	✓	A	95	74.15	74.37
WavLM-Plus [9]	✓	A	95	75.28	75.36
SVFAP [61]	✓	V	78	75.15	75.01
MAE-DFER [58]	✓	V	85	75.91	75.56
VQ-MAE-AV [54]	✓	A+V	30	–	84.80
HiCMAE [60]	✓	A+V	81	**87.96**	**87.99**
AVF-MAE++	✓	A+V	512	87.44	87.57

(j) IEMOCAP

Method	SSL	Mod.	#PS	UAR	WAR
Wav2vec 2.0 [1]	✓	A	95	69.88	67.32
HiCMAE [60]	✓	A+V	81	68.21	68.36
AVBERT [35]	✓	V	37	–	45.80
AVF-MAE++ (L)	✓	A+V	303	69.86	71.65
AVF-MAE++ (H)	✓	A+V	512	**72.71**	**73.83**

(k) AVCAffe

Method	SSL	Mod.	#PS	Arousal	Valence
VGG + MC3 [55]	×	A+V	47	38.90	41.70
HiCMAE [60]	✓	A+V	81	43.18	44.20
AVF-MAE++ (B)*	✓	A+V	169	43.02	46.93
AVF-MAE++ (L)*	✓	A+V	303	45.21	47.83
AVF-MAE++ (H)*	✓	A+V	521	**47.25**	**49.66**

(l) Werewolf-XL

Method	SSL	Arousal	Valence	Dominance
eGeMAPS [19]	×	23.45	8.08	31.15
VGGFace [51]	×	7.24	62.96	14.30
SVFAP [61]	✓	23.51	67.11	34.61
HiCMAE [60]	✓	33.74	69.23	40.66
AVF-MAE++*	✓	**44.99**	**72.19**	**52.35**

Table 2. Performance comparisons of AVF-MAE++ with state-of-the-art CEA and DEA methods on twelve datasets. Mod.: Modality. #PS: Parameters in millions. A: Audio. V: Video. A+V: Audio + Video. *: The results are obtained without progressive training since the disalignment of label semantics. –: Unavailable results. We highlight the best performance in **bold** and underline the second performance.

Method	SAMM	CASME II	SMIC	CAS(ME)^3	MMEW
STSTNet [40]	65.88	83.82	68.01	37.95	80.37
μ-BERT [48]	–	90.34	**85.50**	56.04	–
CapsuleNet [65]	62.09	70.68	58.20	–	67.62
EMR [41]	77.54	82.93	74.61	36.13	81.49
RCN-A [74]	76.01	85.12	63.26	39.28	–
LBP-TOP [83]	39.54	70.26	20.00	21.78	64.23
HTNet [70]	81.31	**95.32**	80.49	57.67	**84.33**
FeatRef [87]	73.72	89.15	70.11	34.93	82.11
AVF-MAE++ (B)	81.58	93.58	83.23	63.18	83.76
AVF-MAE++ (L)	**82.53**	94.03	83.79	**67.88**	83.41
AVF-MAE++ (H)	81.62	94.11	83.55	65.34	**84.33**

Table 3. Performance comparisons of AVF-MAE++ and state-of-the-art MER methods in terms of UF1 (%) on five datasets.

cross-modal features to boost performance. For instance, AVF-MAE++ exceeds UniLearn [11], which pre-trains on both images and videos, by 2.33% UAR and 1.80% WAR on MAFW (11-class). (3) As the capacity of AVF-MAE++ increases, the performance gains from Base to Large are steadily obvious across all the datasets. However, the gains from Large to Huge are much smaller on certain datasets, aligning with the trends in general vision domains [67, 75]. (4) Despite the PSI strategy's efforts to mitigate overfitting, performance still declines slightly on smaller target datasets (*e.g.*, RAVDESS [44]), indicating that large models are particularly prone to overfitting on limited tuning data, which remains a crucial challenge for further improvements.

DEA. We follow the analysis pipeline of HiCMAE to conduct comparisons with previous methods on two datasets, as shown in Tab. 2. It can be clearly seen that AVF-MAE++ outperforms baselines by large margins. Specifically, AVF-MAE++ (H) exceeds the previous best results by 4.07% WA-F1 in Arousal and 5.46% WA-F1 in Valence on AVCAffe [55]. Besides, our method exhibits the largest gain of 11.69% PCC across dimensions on Werewolf-XL [79].

MER. To verify the general applicability of AVF-MAE++, we further evaluate it on the MER task. Different from the above two tasks, MER datasets generally lack audio inputs. We thus only utilize the pre-trained video encoder to con-

Method	Time	Speedup	#PS	MAFW	MER24
HiCMAE [60]	115.45h	–	99	56.17	70.95
Dual masking + Vanilla LGI-Former	77.45h	**1.49×**	142	55.11	69.42
Dual masking + Improved LGI-Former	79.07h	**1.46×**	163	56.12	70.36

Table 4. Ablation comparisons of our dual masking & modality encoder (*i.e.*, Improved LGI-Former) with HiCMAE [60]. We only report results in terms of WAR (%). MER24: MER24-T&V.

DiER Units	HAFE Layer	MAFW		MER24-T&V		IEMOCAP	
		UAR	WAR	WAR	WA-F1	UAR	WAR
×	×	41.89	56.31	70.21	69.62	67.46	69.33
✓	×	42.58	56.81	70.97	70.26	68.56	70.02
×	✓	42.55	56.77	70.65	70.13	68.63	69.86
✓	✓	**42.96**	**57.02**	**71.40**	**70.72**	**68.86**	**70.45**

Table 5. Ablation study on the components of IAV-CL Module.

struct the overall training pipeline on five datasets. As illustrated in Tab. 3, AVF-MAE++ achieves competitive results, exhibiting the largest improvement of 10.21% UF1 compared to HTNet [70] on CAS(ME)^3 [36]. Moreover, we find that on certrain datasets, there exist sharper performance declines when scaling model from Large to Huge, impressing the overfitting conclusions drawn from CEA task.

4.3. Ablation Studies

To investigate the crucial design factors of AVF-MAE++, we systematically conduct in-depth ablation studies on MAFW (11-class) [42] and MER24-T&V [39] datasets.

Impact of Dual Masking & Improved Modality Encoder. Tab. 4 presents the influence of our audio-visual dual masking strategy and improved modality encoder on AVFA performance. Specifically, we employ AVF-MAE++ (B) to fairly compare with the audio-visual encoder-only masking strategy and vanilla ViT [17] of HiCMAE-B [60] on the VoxCeleb2-dev [15] pre-training dataset utilizing 100 epochs. We determine that both our new dual masking strategy and the introduced modality encoder can make positive difference on computational efficiency, exhibiting 1.46× speedup with competitive outcomes.

Evaluation on components of IAV-CL Module. We evaluate the effectiveness of the DiER Units and HAFE Layer in IAV-CL Module using AVF-MAE++ (B), as displayed in Tab. 5. We conduct pre-training on the built hybrid dataset and further fine-tune models on IEMOCAP [4]. Note that when conducting ablative test on HAFE Layer, we deploy the original fusion modules of HiCMAE [60] to construct the hierarchical integration manner. From Tab. 5, we conclude that the coupling use of DiER Units and HAFE Layer leads to the highest improvement, indicating their effectiveness in correlations capture of intra- and inter-modalities.

Ablation Study on the Number of DiER Units. To determine the optimal stacked number of the DiER Units, we conduct ablation studies at different number using AVF-MAE++ (L), as presented in Tab. 6. The results indicate that the stacked number is not directly proportional to performance gains, as too many units lead to overly dense interactions, resulting in increased complexity and instability.

Stacked Number	= 1		= 2		= 4	
	UAR	WAR	UAR	WAR	UAR	WAR
MAFW	43.07	56.83	**43.22**	**57.69**	43.14	57.05
MER24-T&V	61.57	69.98	**62.46**	**71.09**	62.11	70.83

Table 6. Ablation study on the stacked number of DiER Units.

Method	Pre-training Dataset	MAFW		MER24-T&V	
		UAR	WAR	WAR	WA-F1
AVF-MAE++ (B)	VoxCeleb2-dev	42.80	56.64	70.93	70.51
AVF-MAE++ (B)	Unlabeled Hybrid	**42.96**	**57.02**	**71.40**	**70.72**
Δ*Metrics*	–	**+ 0.16%**	**+ 0.38%**	**+ 0.47%**	**+ 0.21%**
AVF-MAE++ (L)	VoxCeleb2-dev	42.67	57.01	70.21	69.02
AVF-MAE++ (L)	Unlabeled Hybrid	**43.22**	**57.69**	**71.09**	**70.32**
Δ*Metrics*	–	**+ 0.55%**	**+ 0.68%**	**+ 0.88%**	**+ 1.30%**
AVF-MAE++ (H)	VoxCeleb2-dev	43.59	57.22	70.45	69.42
AVF-MAE++ (H)	Unlabeled Hybrid	**44.02**	**57.79**	**71.23**	**70.41**
Δ*Metrics*	–	**+ 0.43%**	**+ 0.57%**	**+ 0.78%**	**+ 0.99%**

Table 7. Ablation comparisons on the pre-training data scaling.

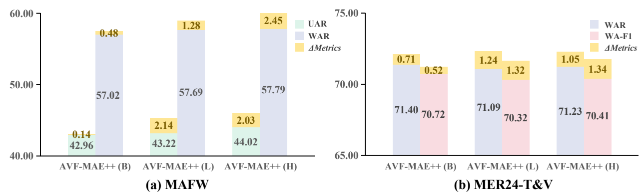

Figure 3. Ablation explorations on the progressive training.

Effectiveness on Data Scaling. As shown in Tab. 7, we assess the effects of pre-training data scaling on AVF-MAE++ using VoxCeleb2-dev [15] and our unlabeled hybrid dataset. We figure out that data scaling consistently boosts performance across all the metrics, emphasizing the importance of data size and diversity for AVFA mask autoencoding.

Contribution of the PSI Strategy. We investigate the contribution of our introduced PSI strategy, as illustrated in Fig. 3. The outcomes indicate that AVF-MAE++ demonstrates superior performance, highlighting its effectiveness in smooth adaptation from pre-training to fine-tuning.

5. Conclusion and Discussions

In this paper, we aim to investigate the scaling properties of audio-visual MAE for AVFA. Thanks to our core designs of dual masking strategy, model architecture, and progressive training pipeline, we are able to successfully train the first hundred-million audio-visual MAE denoted AVF-MAE++ on the currently largest AVFA pre-training dataset. Extensive experiments across 17 datasets verify the superiority of the AVF-MAE++. Our work emphasizes that audio-visual masked autoencoders are scalable and general AVFA represention learners. We hope this work can serve as a foundation and inspire more research on AVFA pre-training.

Despite promising results, challenges persist. Overfitting on small datasets remains a clear bottleneck even with our PSI strategy, and performance seems to saturate on certain datasets as the model capacity grows. Moreover, our data scaling is limited compared to general vision domains [75], leaving pre-training on amplified AVFA data unexplored. We focus on tackling these challenges in the future plans.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 7

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 6

[3] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5826–5846, 2021. 6

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008. 6, 8

[5] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2016. 6

[6] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 6

[7] Chen Chen, Dong Wang, and Thomas Fang Zheng. Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6

[8] Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. *arXiv preprint arXiv:2407.02157*, 2024. 7

[9] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16 (6):1505–1518, 2022. 7

[10] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*, 2024. 7

[11] Yin Chen, Jia Li, Yu Zhang, Zhenzhen Hu, Shiguang Shan, Meng Wang, and Richang Hong. Unilearn: Enhancing dynamic facial expression recognition through unified pre-training and fine-tuning on images and videos. *arXiv preprint arXiv:2409.06154*, 2024. 7

[12] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3884–3892, 2020. 5

[13] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint arXiv:2406.11161*, 2024. 2

[14] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4673–4682, 2024. 7

[15] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 6, 8

[16] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9(1):116–129, 2016. 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3, 4, 6, 8

[18] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 6

[19] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015. 7

[20] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang. Cn-celeb: a challenging chinese speaker recognition dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7604–7608. IEEE, 2020. 6

[21] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 3

[22] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023. 3

[23] Lucas Goncalves, Seong-Gyun Leem, Wei-Cheng Lin, Berrak Sisman, and Carlos Busso. Versatile audio-visual learning for emotion recognition. *IEEE Transactions on Affective Computing*, 2024. 2

[24] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. 3

[25] Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. Cross-mae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26721–26731, 2024. 3

[26] Qiu Han, Gongjie Zhang, Jiaxing Huang, Peng Gao, Zhang Wei, and Shijian Lu. Efficient mae towards large-scale vision transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 606–615, 2024. 3

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[28] 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. 1, 3

[29] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019. 1

[30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460, 2021. 2, 7

[31] Yuchen Hu, Ruizhe Li, Chen Chen, Heqing Zou, Qiushi Zhu, and Eng Siong Chng. Cross-modal global interaction and local alignment for audio-visual speech recognition. *arXiv preprint arXiv:2305.09212*, 2023. 5

[32] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35: 28708–28720, 2022. 4

[33] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[34] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020. 6

[35] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. 7

[36] Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2782–2800, 2022. 6, 8

[37] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*, pages 1–6. IEEE, 2013. 6

[38] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614, 2023. 6

[39] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2404.17113*, 2024. 6, 8

[40] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–5. IEEE, 2019. 7

[41] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–4. IEEE, 2019. 7

[42] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 24–32, 2022. 2, 6, 7, 8

[43] 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. 2

[44] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 6, 7

[45] Yaxiong Ma, Yixue Hao, Min Chen, Jincai Chen, Ping Lu, and Andrej Košir. Audio-visual emotion fusion (avef): A deep efficient weighted approach. *Information Fusion*, 46: 184–192, 2019. 2

[46] Albert Mehrabian. Communication without words. In *Communication theory*, pages 193–200. Routledge, 2017. 1

[47] Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27186–27196, 2024. 3

[48] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1482–1492, 2023. 7

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[50] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445, 2000. 1

[51] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. 7

[52] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Xiang Wang, Yuehuan Wang, Yiliang Lv, Changxin Gao, and Nong Sang. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*, 26:218–233, 2023. 3

[53] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 7

[54] Samir Sadok. *Audiovisual speech representation learning applied to emotion recognition*. PhD thesis, CentraleSupélec, 2024. 3, 7

[55] Pritam Sarkar, Aaron Posen, and Ali Etemad. Avcaffe: a large scale audio-visual dataset of cognitive load and affect for remote work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 76–85, 2023. 1, 6, 7

[56] Dagmar Schuller and Björn W Schuller. The age of artificial emotional intelligence. *Computer*, 51(9):38–46, 2018. 1

[57] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5484–5494, 2023. 3

[58] 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. 1, 2, 3, 4, 7

[59] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325, 2023. 2

[60] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion*, 108:102382, 2024. 1, 2, 3, 5, 6, 7, 8

[61] Licai Sun, Zheng Lian, Kexin Wang, Yu He, Mingyu Xu, Haiyang Sun, Bin Liu, and Jianhua Tao. Svfap: Self-supervised video facial affect perceiver. *IEEE Transactions on Affective Computing*, 2024. 1, 2, 3, 6, 7

[62] 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. 2, 3, 6, 7

[63] Minh Tran, Yelin Kim, Che-Chun Su, Cheng-Hao Kuo, and Mohammad Soleymani. Saaml: A framework for semi-supervised affective adaptation via metric learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6004–6015, 2023. 7

[64] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, page 6558. NIH Public Access, 2019. 2

[65] Nguyen Van Quang, Jinhee Chun, and Takeshi Tokuyama. Capsulenet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 7

[66] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023. 2

[67] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 1, 2, 3, 4, 6, 7

[68] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2531–2539, 2022. 2

[69] Yifan Wang, Xuecheng Wu, Jia Zhang, Mohan Jing, Keda Lu, Jun Yu, Wen Su, Fang Gao, Qingsong Liu, Jianqing Sun, et al. Building robust video-level deepfake detection via audio-visual local-global interactions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11370–11376, 2024. 2

[70] Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. Htnet for micro-expression recognition. *Neurocomputing*, 602:128196, 2024. 7, 8

[71] Jongbhin Woo, Hyeonggon Ryu, Arda Senocak, and Joon Son Chung. Speech guided masked image modeling for visually grounded speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8361–8365. IEEE, 2024. 3

[72] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, 3:e12, 2014. 1

[73] Xuecheng Wu, Heli Sun, Junxiao Xue, Ruofan Zhai, Xiangyan Kong, Jiayu Nie, and Liang He. emotions: A large-scale dataset for emotion recognition in short videos. *arXiv preprint arXiv:2311.17335*, 2023. 1, 2

[74] Zhaoqiang Xia, Wei Peng, Huai-Qian Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model

and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29: 8590–8605, 2020. 7

[75] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10365–10374, 2023. 3, 7, 8

[76] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014. 6

[77] Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung. Attentive modality hopping mechanism for speech emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3362–3366. IEEE, 2020. 7

[78] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022. 7

[79] Kejun Zhang, Xinda Wu, Xinhang Xie, Xiaoran Zhang, Hui Zhang, Xiaoyu Chen, and Lingyun Sun. Werewolf-xl: A database for identifying spontaneous affect in large competitive group interactions. *IEEE Transactions on Affective Computing*, 14(2):1201–1214, 2021. 1, 6, 7

[80] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE transactions on circuits and systems for video technology*, 28(10): 3030–3043, 2017. 1

[81] Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2, 7

[82] Zhicheng Zhang, Pancheng Zhao, Eunil Park, and Jufeng Yang. Mart: Masked affective representation learning via masked temporal distribution distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12830–12840, 2024. 2

[83] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007. 7

[84] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 7

[85] Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023. 7

[86] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021. 7

[87] Ling Zhou, Qirong Mao, Xiaohua Huang, Feifei Zhang, and Zhihong Zhang. Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognition*, 122:108275, 2022. 7

[88] 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. 6

[89] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2