Hierarchical Quantized Cross-modal Masked Autoencoders for Audio-Visual Emotion Recognition

Zeyang Zhang
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
zeyangz@andrew.cmu.edu

Carlos Busso
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
busso@cmu.edu

Abstract

Cross-modal self-supervised learning has shown strong potential for learning representations for *audio-visual emotion recognition* (AVER). However, most existing methods operate entirely in continuous embedding spaces, which can inadvertently preserve irrelevant patterns, spurious correlations, and modality-specific noise. To improve robustness and generalization of AVER models under noisy conditions and domain shift, we propose the *hierarchical quantized cross-modal masked autoencoders* (HQC-MAE) framework. This method integrates discrete representation learning into masked data modeling and contrastive learning via learned codebooks, creating hierarchical information bottlenecks in a multimodal setting. We perform cross-modal contrastive learning in a discrete space, encouraging semantically aligned audio and video samples to match in a structured embedding space. Our vector quantization modules are integrated into pretrained unimodal foundation models (VideoMAEv2 for video and AudioMAE++ for audio), enabling us to leverage large-scale pretrained unimodal models rather than training them from scratch. Experiments on emotion recognition benchmarks across diverse settings show that HQC-MAE consistently outperforms AVER baselines using continuous representations.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

Multimodal Machine Learning, Self-Supervised Learning, Masked Autoencoders, Vector Quantization, Contrastive Learning, Multimodal Representation

ACM Reference Format:
Zeyang Zhang and Carlos Busso. 2026. Hierarchical Quantized Cross-modal Masked Autoencoders for Audio-Visual Emotion Recognition. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3776574.3831121

1 Introduction

Audio-visual emotion recognition (AVER) is a critical problem in affective computing with a variety of applications in human-computer interaction, robotics, education, and healthcare. The system must infer the affective state of people from high-dimensional and noisy signals [8, 38, 41, 44]. Early studies and modern machine learning models consistently report that fusing audio and video improves accuracy and robustness for emotion recognition task compared to using either modality alone [9, 11, 41, 53, 65]. Recently, *self-supervised learning* (SSL) has become an appealing way for learning representations for AVER under scenarios with limited data [5, 13, 27, 28]. However, most SSL frameworks are vulnerable to shortcut learning: high-capacity models may rely on easy-to-use cues that correlate with labels in the training data rather than the underlying affective signal [18]. This issue is amplified in multimodal pretraining because each modality contains distinct, non-corresponding noise sources, such as illumination changes, camera viewpoint, and background motion in videos, and room acoustics, reverberation, and channel effects in audio [10, 43]. These noises do not reliably align across modalities, but can still be encoded in continuous embeddings. Therefore, the learned representations can be contaminated by modality-specific noise and spurious correlations, reducing their transferability and robustness for downstream tasks [64]. Our vision is that performing SSL in a discrete embedding space can strengthen cross-modal alignment and improve model robustness under noisy conditions and domain shifts.

Two families of multimodal SSL objectives are particularly influential: (i) contrastive learning, such as image-text (e.g., CLIP, ALBEF, ALIGN), audio-text (e.g., CLAP, AudioCLIP), and video-text (e.g., VideoCLIP, VATT), and (ii) masked data modeling, such as MultiMAE and FLAVA [1, 4, 16, 25, 35, 39, 46, 48, 60, 62]. For audio-visual SSL specifically, a popular approach is to combine masked data modeling with cross-modal contrastive learning objectives to learn general-purpose joint representations, such as Audiovisual *masked autoencoders* (MAEs), CAV-MAE, and MAViL [19, 24, 30]. Building on these general-purpose learners, several studies adapt them for AVER and get promising results, such as HiCMAE [50] and AVF-MAE++ [59]. Despite these gains, all of these frameworks still operate entirely in continuous embedding spaces, making the models more vulnerable to noise rather than learning robust, semantically aligned representations.

This study proposes the *hierarchical quantized cross-modal masked autoencoders* (HQC-MAE) framework. The key innovation of this proposed method is applying *vector quantization* (VQ) as discrete information bottlenecks in contrastive learning and masked data modeling. These bottlenecks discretize representations at multiple layers, acting as a regularizer that suppresses modality-specific noise while encouraging robust, structured, and partially disentangled representations. HQC-MAE performs contrastive learning using compositional discrete representations: each patch token selects a discrete code, and sample-level features are formed by

ICMI '26, Napoli, Italy

ACM ISBN 979-8-4007-2318-6/26/10
https://doi.org/10.1145/3776574.3831121

composing (e.g., averaging) the selected codebook vectors. This token-level discretization forces each audio-visual patch to map into a semantic primitive, filtering local nuisance factors in video and audio, and enabling implicit semantic clustering that facilitates cross-modal alignment. To leverage recent advances in unimodal masked autoencoders, HQC-MAE integrates pretrained encoders (VideoMAEv2 for video and AudioMAE++ for audio) rather than pretraining all components from scratch [55, 63]. We use Gumbel-Softmax with temperature annealing and a diversity loss to avoid index collapse in VQ, and achieve a smooth convergence [17, 54].

To the best our knowledge, HQC-MAE is the first self-supervised AVER model that integrates VQ at intermediate layers of multimodal MAEs, performing cross-modal contrastive learning in a discrete embedding space. For evaluation, we finetune the pretrained model on AVER benchmarks and apply chunk-level representation aggregation during inference time [42]. We report performance using *weighted average recall* (WAR) and *unweighted average recall* (UAR). On the MSP-IMPROV corpus, HQC-MAE outperforms strong continuous baselines, achieving +2.43 WAR / +7.30 UAR over HiCMAE, and surpassing the state-of-the-art AVF-MAE++ by +2.27 WAR / +0.94 UAR [10, 15, 50, 59]. We also conduct ablation studies to validate the contribution of quantization in the encoders, the fusion stage, and the contrastive objective. In summary, our contributions are:

- We introduce a cross-modal MAE framework that uses multiple learnable VQ codebooks at intermediate layers of audio, video, and fusion encoders to create hierarchical discrete bottlenecks.
- We propose cross-modal contrastive learning on compositional discrete representations, encouraging structured cross-modal alignment in a discrete geometry to improve generalization ability and robustness under noisy conditions.
- We achieved consistent gains over continuous alternatives and validate the importance of the proposed VQ modules.

2 Related Works

Masked Autoencoders: The MAE is a self-supervised learning approach that reconstructs heavily masked patches with an asymmetric encoder-decoder design [27]. The resulting pretrained backbones transfer well to diverse downstream vision tasks, including image classification, object detection, instance segmentation, and semantic segmentation [26, 40, 49]. This MAE design has been successfully adapted beyond images. AudioMAE [31] applies MAE-style pretraining to audio by masking and reconstructing mel-spectrogram patches. Recent work such as AudioMAE++ [63] shows that modern transformer components can further improve masked audio representation learning. For video, VideoMAE adopts tube masking over spatiotemporal tokens for efficient video pretraining, while VideoMAEv2 scales masked video modeling with dual masking [51, 55]. Our method builds on AudioMAE++ and VideoMAEv2, using their pretrained weights for initialization.

Contrastive Audio-Visual Learning: Contrastive audio-visual learning is a self-supervised approach that leverages the natural temporal alignment of audio and video to learn joint representations by matching corresponding audio-visual segments and contrasting mismatched pairs [37]. Early work framed this supervision as correspondence classification [3]. Later methods explored alternative cross-modal objectives such as cross-modal clustering [2]. More recent approaches use contrastive instance discrimination explicitly. *Audio-visual instance discrimination* (AVID) [45] trains audio and video encoders by treating temporally aligned audio-visual segments as positive pairs and mismatched segments as negatives. *Video-audio-text transformer* (VATT) [1] demonstrates that Transformer-based encoders can be trained end-to-end with multimodal contrastive objectives to learn scalable semantic representations directly from raw video, audio, and text. Audio-visual contrastive learning with temporal self-supervision [33] strengthens audio-visual contrastive training by adding temporal auxiliary tasks, such as recognizing playback speed or temporal direction. *Sequential contrastive audio-visual learning* (SCAV) [52] moves beyond instance level features and performs contrastive learning over non-aggregated sequences using sequential distances. However, these methods typically contrast continuous audio-visual embeddings directly. Our work instead performs contrastive learning on compositional discrete representations.

Audio-Visual Emotion Recognition: AVER aims to predict human emotion from synchronized speech and facial video [20, 21, 23]. Recently, self-supervised audio-visual representation learning has become a major direction in AVER. For example, VQ-MAE-AV [47] is a vector quantized masked autoencoder designed for audiovisual speech representation learning. It first learns VQ-VAE tokenizers to convert raw audio and visual streams into discrete tokens, then trains a multimodal MAEs to reconstruct masked tokens, and finally finetunes the pretrained model for emotion recognition. *Hierarchical contrastive masked autoencoder* (HiCMAE) [50] introduces hierarchical cross-modal contrastive learning in MAEs with hierarchical skip connections to explicitly encourage emotion-relevant features at multiple depths. AVF-MAE++ [59] further investigates how to scale MAE-style audio-visual pretraining for affective facial analysis, proposing architectural and training refinements such as dual masking, more efficient modality encoders, and explicit correlation-learning components. In addition to pretraining a unified audio-visual encoder, VAEmo [14] adopts a two-stage training scheme that further injects emotion knowledge by aligning audio-visual features with affective descriptions generated by multimodal *large language models* (LLMs). Despite this progress, most MAE-based AVER frameworks still learn and align continuous embeddings. Our HQC-MAE is designed to fill this gap with discrete representation learning. The most related work to ours is VQ-MAE-AV, but this approach requires a two-stage training and only applies quantization at the input token level [47].

Discrete Representation Learning: Discrete Representation learning aims to map continuous signals into a finite set of discrete units (a codebook) so that the model can reason, compress, or generate using token-like representations rather than dense vectors. A common approach is VQ, where an encoder produces a continuous latent and each latent vector is assigned to a codebook entry, creating a discrete bottleneck that supports efficient storage and scalable sequence modeling. VQ has been highly successful in unimodal settings. In vision, discrete latents enable transformer-based generation to create high fidelity images (e.g., VQ-VAE and VQGAN) [54]. In speech/audio, discrete units are widely used in self-supervised pretraining (e.g., vq-wav2vec learns quantized speech units) and

in token-based generation or codec pipelines (e.g., BigCodec, TS3-Codec, WavTokenizer) [6, 34, 58, 61]. While discrete tokenizers and VQ objectives are now well developed for unimodal models, less explored directions are how to apply VQ inside multimodal representation learning pipelines and prevent the issue of index collapse under cross-modal objectives.

3 Method

This section describes our proposed HQC-MAE for learning noise-robust representations for AVER. Fig. 1 shows the pretraining framework, which has three main blocks: (i) pretrained modality-specific encoders, (ii) cross-modal fusion encoder, (iii) decoders. Fig. 2 shows the finetuning framework during inferences where we discard the decoders, using only the audio encoder, video encoder, and cross-modal fusion encoder to produce emotion predictions.

3.1 Pretrained Modality-Specific Encoders

We initialize both audio and video encoders with pretrained unimodal models instead of training them from scratch. For video, we use VideoMAEv2-Base, a 12-layer Vision Transformer with 768-dimensional embeddings pretrained on video frames [55]. For audio, we utilize AudioMAE++ Tiny, a 12-layer encoder with 192-dimensional embeddings pretrained on audio spectrograms [63]. Both encoders use their native input resolutions: 224×224 RGB frames for video with tubelet size 2 and patch size 16x16, and 200×80 mel-spectrograms for audio with patch size 4×16.

Each encoder extracts intermediate features at layers 3, 7, and 11 in addition to the final layer output. All features are projected to a shared d-dimensional space ($d = 512$) through learned linear projection layers. This hierarchical feature extraction framework enables multi-scale representation learning, where early layers capture low-level details and deeper layers encode high-level semantics.

3.2 Vector Quantization Modules

A main contribution of our method is the introduction of vector quantization modules in a multimodal setting that discretize representations at multiple depths within both audio and video encoders and after fusion encoders. Our architecture uses 8 learnable codebooks in total:

- Encoder-level codebooks: 3 codebooks in the video encoder and 3 codebooks in the audio encoder are applied directly after layers 3, 7, 11 to enable hierarchical discrete representation learning.
- Fusion-level codebooks: 1 codebook for the audio branch and 1 codebook for the video branch are applied after the cross-modal fusion encoder to produce the final quantized representations (PFVQ in Fig. 1).

Each codebook uses the Gumbel-Softmax [32] quantization with gradient optimization, allowing end-to-end differentiable training. Given an input feature sequence $\mathbf{z} \in \mathbb{R}^{N\times d}$, we compute logits via a learned projection $\mathbf{W} \in \mathbb{R}^{d\times K}$, add Gumbel noise, and apply softmax to obtain soft code assignments:

$$\mathbf{z}_q = \text{softmax}\left(\frac{\mathbf{zW} + \mathbf{g}}{\tau}\right)\mathbf{C}, \tag{1}$$

where $\mathbf{g} \sim \text{Gumbel}(0, 1)$ is Gumbel noise sampled during training, τ is the temperature parameter, and $\mathbf{C} \in \mathbb{R}^{K\times d}$ is the learnable codebook with K codes. During training, the forward pass uses hard assignments while gradients flow through soft assignments.

We apply temperature annealing, starting at $\tau = 2.0$ and exponentially decaying with a factor 0.999995 per training step to a minimum of $\tau = 0.5$. High temperature in early training promotes broader codebook exploration and helps mitigate index collapse, while lower temperature in later stages enforces stronger discrete assignments and stabilizes code selection.

To encourage uniform codebook utilization and prevent index collapse, we add *Kullback-Leibler* (KL) divergence regularization term for each codebook:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{K} p_i \log\left(p_i \cdot K\right), \tag{2}$$

where $p_i = \frac{1}{N}\sum_{n=1}^{N} \text{softmax}(\mathbf{z}_n\mathbf{W})_i$ is the average soft assignment probability for code i across all N tokens, $\mathbf{z}_n \in \mathbb{R}^d$ denotes the n-th token in the input feature sequence $\mathbf{z}$, and K is the codebook size.

3.3 Quantized Hierarchical Cross-Modal Contrastive Learning

The main idea explored in this paper is using the quantized features for contrastive learning in a multimodal problem. We expect that this strategy enforces the cross-modal alignment to be expressed through a discrete and compositional vocabulary. It also encourages the codebooks to learn semantically meaningful discrete units that capture audio-visual correspondences, avoiding spurious correlations.

Building on the *hierarchical cross-modal contrastive learning* (HCMCL) objective in HiCMAE [50], our HQC-MAE applies VQ on the continuous intermediate features in modality-specific encoders for cross-modal contrastive learning. For each selected intermediate layer $l \in \{3, 7, 11\}$, we pass the continuous features through the corresponding VQ module to obtain quantized video features $\mathbf{z}_q^{v,l} \in \mathbb{R}^{N_v\times d}$ and audio features $\mathbf{z}_q^{a,l} \in \mathbb{R}^{N_a\times d}$, where N_v and N_a are the number of visible video and audio tokens respectively. These quantized features are then mean-pooled and L2 normalized:

$$\bar{\mathbf{z}}^{v,l} = \text{norm}\left(\frac{1}{N_v}\sum_{i=1}^{N_v}\mathbf{z}_{q,i}^{v,l}\right), \bar{\mathbf{z}}^{a,l} = \text{norm}\left(\frac{1}{N_a}\sum_{i=1}^{N_a}\mathbf{z}_{q,i}^{a,l}\right).$$

Given a batch of size B, we form the similarity matrix $\mathbf{S}^l \in \mathbb{R}^{B\times B}$ with entries

$$S_{ij}^l = \frac{\bar{\mathbf{z}}_i^{v,l} \cdot \bar{\mathbf{z}}_j^{a,l}}{\tau_c}, \tag{3}$$

where $\tau_c = 0.07$ is the contrastive temperature. The InfoNCE loss is computed over the similarity matrix:

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2}\sum_{l\in\{3,7,11\}}\left(\text{CE}(\mathbf{S}^l, \mathbf{y}) + \text{CE}((\mathbf{S}^l)^\top, \mathbf{y})\right), \tag{4}$$

where $\text{CE}(\cdot,\cdot)$ is the cross-entropy loss and $\mathbf{y}$ is the ground truth label vector indicating the matched audio-video pairs (i.e., the diagonal entries of $\mathbf{S}^l$, which correspond to positive pairs).

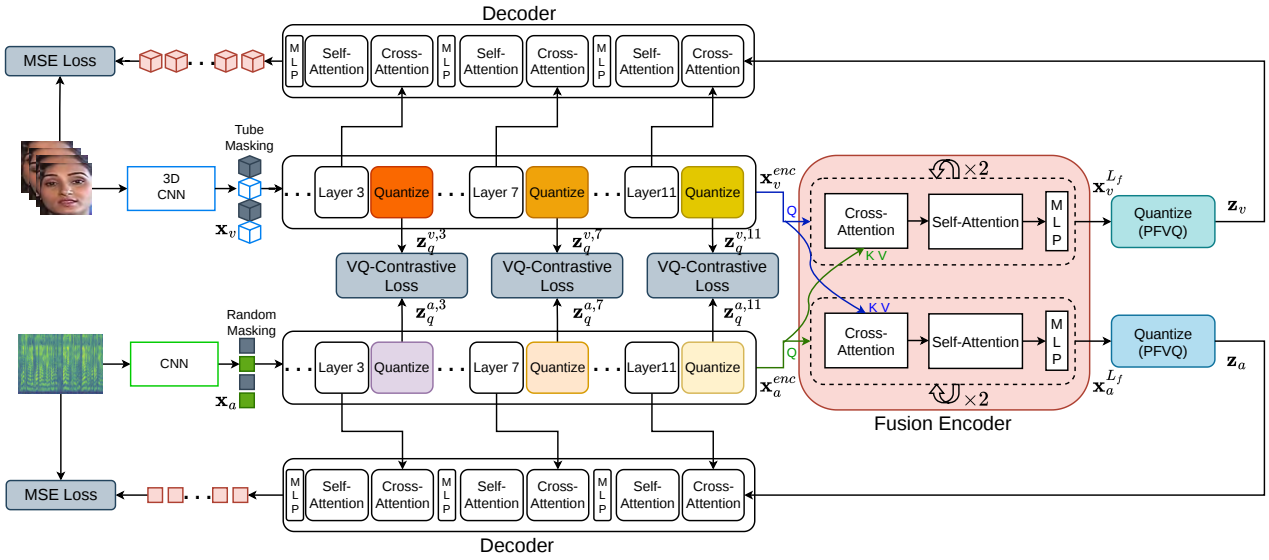

Figure 1: Overview of the HQC-MAE pretraining architecture. The framework introduces vector quantization modules across the video encoder, audio encoder, and cross-modal fusion encoder and performs cross-modal contrastive learning in discrete embedding spaces.

3.4 Gradient Flow Design

A key design choice of our HQC-MAE is to separate the gradient paths for the VQ codebooks in modality-specific encoders. These codebooks are optimized only by the contrastive loss and the KL regularization term. The MAE reconstruction loss does not back-propagate through these codebooks. This strategy prevents gradient competition between reconstruction objective and contrastive objective. In contrast, the codebooks in the cross-modal fusion encoders are optimized by the MAE reconstruction loss since the decoder receives the quantized fusion features.

3.5 Cross-Modal Fusion Encoder

We use a bidirectional cross-attention fusion encoder with two Transformer layers with 12 attention heads per layer. Let $\mathbf{x}_v^{\text{enc}} \in \mathbb{R}^{N_v \times d}$ and $\mathbf{x}_a^{\text{enc}} \in \mathbb{R}^{N_a \times d}$ denote the video and audio feature sequences output from their modality-specific encoders. For each fusion layer, the video branch computes:

$$\tilde{\mathbf{x}}_v = \mathbf{x}_v^{\text{enc}} + \text{CrossAttn}(\mathbf{x}_v^{\text{enc}}, \mathbf{x}_a^{\text{enc}}),$$
$$\hat{\mathbf{x}}_v = \tilde{\mathbf{x}}_v + \text{SelfAttn}(\tilde{\mathbf{x}}_v), \quad \mathbf{x}_v' = \hat{\mathbf{x}}_v + \text{FFN}(\hat{\mathbf{x}}_v),$$

The audio branch is symmetric, with v and a swapped in the above equations. $\text{CrossAttn}(\mathbf{x}_i, \mathbf{x}_j)$ denotes multi-head cross-attention using queries from modality i (first argument), and keys/values from modality j (second argument); $\text{SelfAttn}(\cdot)$ denotes multi-head self-attention; and $\text{FFN}(\cdot)$ is a two-layer feed-forward network with *Gaussian error linear unit* (GELU) activation [29].

3.6 Post-Fusion Vector Quantization (PFVQ)

After fusion, we apply VQ modules to the final outputs of the cross-modal fusion encoders for the video and audio branches, denoted as $\mathbf{x}_v^{L_f}$ and $\mathbf{x}_a^{L_f}$ respectively:

$$\mathbf{z}_v = \text{VQ}_v(\mathbf{x}_v^{L_f}; C_v), \qquad \mathbf{z}_a = \text{VQ}_a(\mathbf{x}_a^{L_f}; C_a), \tag{5}$$

where $C_v \in \mathbb{R}^{K_v \times d}$ and $C_a \in \mathbb{R}^{K_a \times d}$ are the video and audio codebooks with K_v and K_a codes, respectively. The quantized representations $\mathbf{z}_v$ and $\mathbf{z}_a$ are then passed to their respective decoders for masked reconstruction. The motivation of this block is to regularize fused representations with a discrete bottleneck so that cross-modal interactions are encoded more compactly before reconstruction.

3.7 Pretraining Objectives

Masked Autoencoding Loss. We apply masking with a ratio of 90% for video and 80% for audio. The reconstruction targets are patch-normalized video pixels and audio spectrogram values:

$$\mathcal{L}_{\text{MAE}} = \text{MSE}\left(\hat{x}_v^{\text{mask}}, x_v^{\text{mask}}\right) + \text{MSE}\left(\hat{x}_a^{\text{mask}}, x_a^{\text{mask}}\right), \tag{6}$$

where $\hat{\mathbf{x}}_v^{\text{mask}}$ and $\hat{\mathbf{x}}_a^{\text{mask}}$ are the decoder-reconstructed video and audio patches at masked positions, and $\mathbf{x}_v^{\text{mask}}$ and $\mathbf{x}_a^{\text{mask}}$ are the corresponding ground-truth patches.

VQ Regularization. We sum the KL regularization terms across all 8 codebooks:

$$\mathcal{L}_{\text{VQ}} = \sum_{l \in \{3,7,11\}} \left(\mathcal{L}_{\text{KL}}^{v,l} + \mathcal{L}_{\text{KL}}^{a,l}\right) + \mathcal{L}_{\text{KL}}^{v,\text{fusion}} + \mathcal{L}_{\text{KL}}^{a,\text{fusion}}, \tag{7}$$

where $\mathcal{L}_{\text{KL}}^{v,l}$ and $\mathcal{L}_{\text{KL}}^{a,l}$ are the KL regularization losses for the video and audio encoder codebooks at layer l, and $\mathcal{L}_{\text{KL}}^{v,\text{fusion}}$ and $\mathcal{L}_{\text{KL}}^{a,\text{fusion}}$ are the KL regularization losses for the video and audio fusion codebooks. Each KL term is computed with the definition in Eq. 2.

Objective. The overall pretraining objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MAE}} + \lambda_c \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{VQ}} \mathcal{L}_{\text{VQ}}, \tag{8}$$

where $\lambda_c = 0.0025$ is the contrastive loss weight (Eq. 4) and $\lambda_{\text{VQ}} = 0.0005$ is the VQ regularization weight.

3.8 Finetuning for Emotion Recognition

For downstream AVER, we use the learned hierarchical quantized representations.

Feature Extraction. For the selected layers $l \in \{3, 7, 11\}$ in modality-specific encoders, we quantize the intermediate features with the corresponding VQ modules and obtain token sequences $\mathbf{z}_q^{v,l}$ and $\mathbf{z}_q^{a,l}$. We then compute layer-wise features by mean pooling over all tokens:

$$\bar{\mathbf{z}}_q^{v,l} = \frac{1}{N_v} \sum_{i=1}^{N_v} \mathbf{z}_{q,i}^{v,l}, \qquad \bar{\mathbf{z}}_q^{a,l} = \frac{1}{N_a} \sum_{i=1}^{N_a} \mathbf{z}_{q,i}^{a,l}. \tag{9}$$

Fusion-level quantized features are mean-pooled in the same way to obtain $\bar{\mathbf{z}}_q^{v,\text{fusion}}$ and $\bar{\mathbf{z}}_q^{a,\text{fusion}}$. We then use a learnable weighted sum to aggregate the layer-wise encoder features:

$$\mathbf{f}_v = \sum_{l \in \{3,7,11\}} \alpha_l^v \, \bar{\mathbf{z}}_q^{v,l}, \qquad \mathbf{f}_a = \sum_{l \in \{3,7,11\}} \alpha_l^a \, \bar{\mathbf{z}}_q^{a,l}, \tag{10}$$

where $\boldsymbol{\alpha}^v = \text{softmax}(\mathbf{w}^v)$ and $\boldsymbol{\alpha}^a = \text{softmax}(\mathbf{w}^a)$ with learnable $\mathbf{w}^v, \mathbf{w}^a \in \mathbb{R}^3$, initialized uniformly.

Classification. We concatenate the aggregated encoder-level features with the mean-pooled fusion-level features:

$$\mathbf{f} = \left[\mathbf{f}_v;\ \mathbf{f}_a;\ \bar{\mathbf{z}}_v;\ \bar{\mathbf{z}}_a\right] \in \mathbb{R}^{4d}, \tag{11}$$

where $d = 512$, giving a total dimension of 2,048. This is followed by a LayerNorm and a linear classification head for *multimodal emotion recognition* (MMER).

Objective. The finetuning loss is

$$\mathcal{L}_{\text{finetune}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{VQ}}, \tag{12}$$

where $\mathcal{L}_{\text{VQ}}$ is the sum of KL regularization terms from all 8 codebooks (same as in pretraining; Eq. 7). The VQ modules remain trainable during finetuning, using the same temperature annealing schedule as in pretraining.

4 Experimental Settings

4.1 Databases

We use the VoxCeleb2 [15] dataset for pretraining and the MSP-IMPROV [10], MAFW [43], and CREMA-D [11] corpora for finetuning.

VoxCeleb2 [15] is a large-scale speaker recognition dataset containing over 1 million video clips from 6,112 celebrities. For pretraining, we use 1,091,580 video clips in this dataset. All videos are preprocessed to 25 FPS with audio resampled at a sampling rate of 16kHz.

MSP-IMPROV [10] is an audio-visual emotion recognition dataset containing 8,687 video clips with 4 categorical emotions: Anger (A), Happiness (H), Neutral (N), and Sadness (S). We use a *Leave-one-session-out* (LOSO) cross-validation approach with 6 folds. In each fold, we hold out one session as the test set and train on the remaining sessions. From the training sessions, we reserve a small, disjoint subset as the validation set. Each fold contains approximately 6,500 training samples, 1,000 validation samples, and 1,000 test samples.

MAFW [43] is an in-the-wild audio-visual emotion dataset containing 9,156 video clips with 11 categorical emotions. We follow a 5-fold cross-validation with approximately 7,000 training samples, 300 validation samples, and 1,800 test samples per fold.

CREMA-D [11] is an audio-visual emotion recognition dataset containing 7,438 video clips from 91 actors with 6 categorical emotions: Anger (A), Disgust (D), Fear (F), Happiness (H), Neutral (N), and Sadness (S). We perform a 5-fold cross-validation following the standard speaker-independent splits to ensure fair comparison with the baselines, with approximately 5,500 training samples, 400 validation samples, and 1,476 test samples per fold.

These dataset splits follow the setup used in the baselines (Sec. 4.4) to ensure fair comparison.

4.2 Evaluation Metrics

For emotion recognition, we report WAR and UAR. WAR represents overall accuracy weighted by class frequency. UAR is the macro-averaged recall across all classes, which is particularly important for imbalanced datasets. The best model is selected based on the best WAR score on the validation set.

4.3 Training Details

Pretraining. We sample a clip from each input video by randomly selecting a starting timestamp and extracting 16 frames with a temporal stride of 4 frames. Since we resample all VoxCeleb2 videos to 25 FPS, the resulting clip duration is approximately 2.6 seconds. Each frame is resized to 224×224 to match the native input resolution of VideoMAEv2-Base [55]. We extract the audio waveform segment from the same time window as the selected video clip to ensure audio-video alignment. We convert the waveform to an 80-bin log-mel spectrogram and use an input size of 200×80 (time $\times$ mel) to match the input resolution of AudioMAE++ Tiny [63].

We pretrain our model for 200 epochs on VoxCeleb2 using AdamW optimizer with a learning rate of 3×10^{-4}, 5 warmup epochs, and a cosine learning rate schedule. We apply gradient clipping with a max norm of 3.0. We use a batch size of 48 on six L40S GPUs.

Codebook sizes. We use modality-specific codebook sizes in HQC-MAE. By default, each video codebook contains $K_v = 512$ codes and each audio codebook contains $K_a = 320$ codes, since these settings yield the best AVER performance in our experiments (see Table 4).

Finetuning. When we finetune our pretrained model on AVER benchmarks, we use the same input sizes as pretraining. We train for 100 epochs using AdamW optimizer with base learning rate of 1e-3, minimum learning rate of 1e-6, weight decay of 0.05, and 5 warmup epochs, and a cosine learning rate scheduling. We use a batch size of 12 per GPU on two L40S GPUs.

Given the arbitrary duration of the segments and our fixed analysis window for inference, we further adopt a chunk-level representation during inference time, aggregating multiple short segments to get more robust prediction [42]. This approach better capture the temporal structure in the externalization of emotions. Specifically,

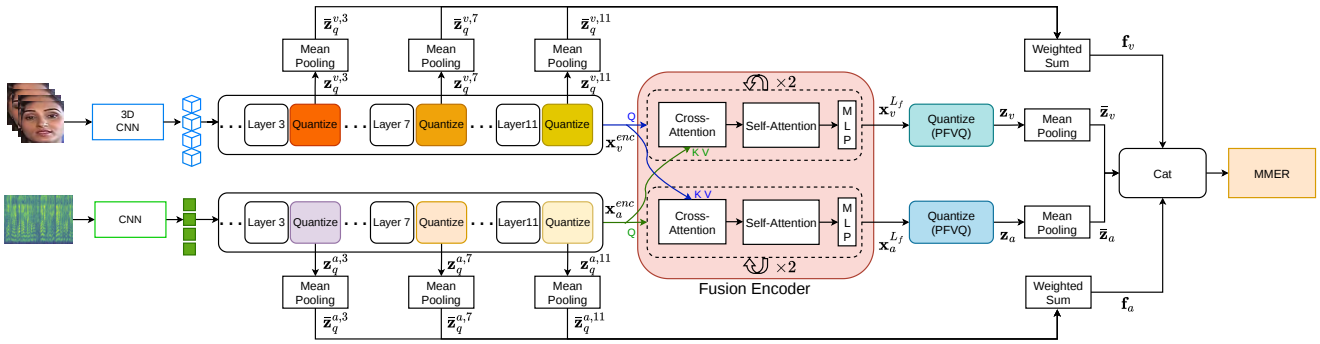

Figure 2: Finetuning architecture for speech emotion recognition. We use quantized vectors to get emotion predictions.

we uniformly sample six temporal segments spanning the entire video and compute the final prediction by averaging the softmax probabilities at the segment-level.

4.4 Baselines

We benchmarked HQC-MAE to compare the performance of our model with three other state-of-the-art self-supervised AVER frameworks: HiCMAE [50], AVF-MAE++ [59], and VAEmo [14]. We choose these models as our baselines because they all combine masked data modeling with contrastive learning for self-supervised audio-visual representation learning.

To highlight the benefit of large-scale self-supervised pretraining, we also include a strong supervised baseline, RAVER [22], which is trained directly on labeled emotion data without a pretraining stage. RAVER has shown better performance than other fully supervised models, so we compare only against RAVER in this category. It uses WavLM [12] for audio encoding and MobileNetV3 [36] for visual encoding with cross-modal attention for multimodal fusion.

We reproduced RAVER and HiCMAE using the same training environment to ensure a fair comparison. We were not able to fully reproduce AVF-MAE++ and VAEmo because their source code or pretrained encoders were unavailable. Nevertheless, comparison with their reported results remains fair and informative because these methods rely on stronger supervision or additional data. AVF-MAE++ adopts a three-stage pipeline, uses datasets beyond VoxCeleb2, and has substantially more parameters. VAEmo further relies on LLM-generated emotion descriptions for knowledge injection, which our method does not use.

5 Results

5.1 Results on AVER Benchmarks

In this section, we evaluate the performance of HQC-MAE on AVER. Each experiment is repeated three times, and we report the average performance. Table 1 summarizes the results on the MAFW, MSP-IMPROV, and CREMA-D datasets.

Comparison with SSL Baselines. As shown in Table 1, HQC-MAE performs consistently well across MSP-IMPROV, CREMA-D, and MAFW datasets, demonstrating that learning discrete representations provides stronger and more transferable features than previous SSL frameworks that rely on only continuous embeddings. HQC-MAE significantly improves performance over HiCMAE [50] on all three datasets (MAFW: +2.53% WAR, +3.27% UAR; MSP-IMPROV: +2.43% WAR, +7.30% UAR; CREMA-D: +1.35% WAR, +1.11% UAR), demonstrating the benefit of introducing VQ as a regularizer to encourage cross-modal alignment and suppress noise. In addition, HQC-MAE achieves comparable performance to AVF-MAE++ [59] while using approximately 74% fewer parameters, suggesting a strong parameter efficiency and leaving room for further gains by scaling up the model. HQC-MAE also consistently outperforms VAEmo [14] on all benchmarks (MAFW: +0.77% WAR, +0.86% UAR; MSP-IMPROV: +1.55% WAR, +4.87% UAR; CREMA-D: +0.33% WAR, +0.18% UAR). While VAEmo is more lightweight than our model, it relies on a two-stage training pipeline and additional multimodal LLMs to generate affective descriptions for knowledge injection.

Comparison with Supervised Learning Baseline. HQC-MAE substantially outperforms the state-of-the-art fully supervised baseline RAVER [22] across all datasets (MAFW: +20.01% WAR, +15.31% UAR; MSP-IMPROV: +2.27% WAR, +1.14% UAR; CREMA-D: +6.75% WAR, +6.55% UAR). These clear improvements demonstrate that self-supervised pretraining on large-scale audio-visual data, combined with discrete cross-modal alignment, provides significantly better initialization than training from scratch using only limited emotion-labeled data.

5.2 Ablation Studies

We study the contribution of VQ in HQC-MAE with a focus on its roles in quantized contrastive learning and cross-modal fusion encoder. In addition to the full model, we consider three variants for ablation studies: (i) HQC-MAE without VQ, (ii) HQC-MAE with *vector quantized contrastive learning* (VQCL) only (i.e., VQ in the encoders), and (iii) HQC-MAE with *post-fusion vector quantization* (PFVQ) only. All variants are pretrained on the Voxceleb2 dataset and finetuned on the MSP-IMPROV corpus with the same hyperparameters and procedures.

Table 1: Performance comparison on categorical emotion recognition datasets MAFW, MSP-IMPROV, and CREMA-D. * denotes results reproduced by us. Through out the paper, we highlight the best result in Bold.

				MAFW		MSP-IMPROV		CREMA-D	
Method	SSL	Mod.	#Params(M)	UAR	WAR	UAR	WAR	UAR	WAR
AVF-MAE++ (Wu et al., 2025)	✓	A+V	521	46.05	**60.24**	70.05	76.07	**86.02**	85.95
VAEmo (Cheng et al., 2025)	✓	A+V	39	45.67	58.91	66.12	76.79	85.71	85.68
RAVER* (Goncalves et al., 2026)	×	A+V	348	31.22	39.67	69.85	76.07	79.34	79.26
HiCMAE* (Sun et al., 2024)	✓	A+V	81	43.26	57.15	63.69	75.91	84.78	84.66
HQC-MAE (ours)	✓	A+V	136	**46.53**	59.68	**70.99**	**78.34**	85.89	**86.01**

Table 2: Ablation study on vector quantization modules evaluated on MSP-IMPROV. VQ: Vector Quantization, VQCL: Vector Quantized Contrastive Learning, PFVQ: Post-Fusion VQ.

VQ	VQCL	PFVQ	UAR	WAR
×	×	×	66.76	76.38
✓	✓	×	70.75	76.72
✓	×	✓	68.93	76.46
✓	✓	✓	**70.99**	**78.34**

Table 3: Ablation study on loss weights. For λ_c, λ_{VQ} is fixed at 0.0005. For λ_{VQ}, λ_c is fixed at 0.0025.

Setting	UAR	WAR
$\lambda_c = 0.001$	69.24	76.25
$\lambda_c = 0.0025$	**70.01**	**77.36**
$\lambda_c = 0.01$	68.41	75.57
$\lambda_{\text{VQ}} = 0.0001$	69.53	77.18
$\lambda_{\text{VQ}} = 0.0005$	**70.01**	**77.36**
$\lambda_{\text{VQ}} = 0.001$	69.50	76.11

As shown in Table 2, removing VQ modules causes a clear degradation. Compared to HQC-MAE without VQ, our full HQC-MAE achieves a +1.96% improvement in WAR and +4.23% improvement in UAR. Applying VQ to only one component (either VQCL or PFVQ) yields smaller gains compared to HQC-MAE without VQ, demonstrating that applying VQ to both hierarchical contrastive learning and cross-modal fusion encoder provides complementary benefits. Specifically, VQ in the contrastive objective enables stronger cross-modal alignment by creating discrete cluster centers in the embedding space. Moreover, VQ in the cross-modal fusion encoder regularizes the fused representations and mitigates overfitting. Therefore, the combination of both mechanisms produces more structured and discriminative emotion representations.

Ablation Study on Loss Weights. We study the effect of the two loss weights in the pretraining objective (Eq. 8): the contrastive loss weight λ_c and the VQ regularization loss weight λ_{VQ}. For these ablations, all models are finetuned on MSP-IMPROV, and UAR/WAR are reported on the MSP-IMPROV validation set.

As shown in Table 3, $\lambda_c = 0.0025$ gives the best validation performance when λ_{VQ} is fixed at 0.0005. A smaller value ($\lambda_c = 0.001$) weakens the cross-modal alignment signal, while a larger value ($\lambda_c = 0.01$) over-emphasizes contrastive learning and hurts reconstruction. When fixing $\lambda_c = 0.0025$, the best performance is also achieved with $\lambda_{\text{VQ}} = 0.0005$. A smaller value ($\lambda_{\text{VQ}} = 0.0001$) provides weaker regularization and may reduce codebook diversity, whereas a larger value ($\lambda_{\text{VQ}} = 0.001$) over-regularizes the codebooks and limits their representational capacity. Thus, we use $\lambda_c = 0.0025$ and $\lambda_{\text{VQ}} = 0.0005$ as the default settings.

Table 4: Effect of codebook size on MSP-IMPROV. These results are reported on the validation set.

K_v	K_a	UAR	WAR
256	160	69.14	75.88
512	320	**70.01**	**77.36**
1024	512	69.55	76.83
2048	1024	68.72	77.01

5.3 Effect of codebook sizes

We analyze how the codebook sizes for video and audio affect downstream AVER on the validation set. As shown in Table 4, the model performance is not monotonic with respect to video and audio codebook sizes (K_v and K_a). An intermediate codebook setting (K_v=512, K_a=320) achieves the best results (77.36% WAR, 70.01% UAR). Reducing and increasing codebook sizes both degrade model performance, suggesting that a moderately compact discrete embedding space provides a better trade-off between representational capacity and regularization. This trend is consistent with previous findings that VQ can act as regularizers by constraining latents and encouraging semantically similar inputs to share codes, while overly large codebooks can weaken this constraint and hurt generalization [56, 57]. Moreover, our audio codebook size (K_a=320) is in line with common choices in speech quantization [7]. Therefore, we use K_v=512 and K_a=320 as a default setting in HQC-MAE.

5.4 Robustness under Noisy Test Conditions

To examine whether introducing VQ into contrastive learning improves robustness, we evaluate the HQC-MAE (VQCL) variant under noisy test conditions by keeping the training and validation sets clean and adding Gaussian noise only to the MSP-IMPROV test set at inference time. We consider two noise settings: $\sigma_v = 0.1$, $\sigma_a = 0.02$ and $\sigma_v = 0.4$, $\sigma_a = 0.2$. This setting is more challenging than standard within-corpus evaluation because the model must recognize emotions from corrupted audio-visual inputs.

As shown in Table 5, HQC-MAE (VQCL) achieves the best performance under noisy test conditions. Compared with HiCMAE, HQC-MAE improves UAR/WAR by +7.14%/+1.84% under moderate

Table 5: Performance comparison on MSP-IMPROV under different Gaussian-noise conditions.

				$\sigma_v = 0.10$ $\sigma_a = 0.02$		$\sigma_v = 0.4$ $\sigma_a = 0.2$	
Method	SSL	Mod.	#Params(M)	UAR	WAR	UAR	WAR
RAVER [22]	×	A+V	348	68.98	73.80	60.11	62.41
HiCMAE [50]	✓	A+V	81	64.33	75.80	51.77	71.38
HQC-MAE (VQCL)	✓	A+V	136	**71.47**	**77.64**	**63.39**	**76.64**

Table 6: Cross-corpus evaluation with training on MSP-IMPROV and testing on CREMA-D.

Method	SSL	Mod.	#Params(M)	UAR	WAR
RAVER [22]	×	A+V	348	66.12	65.55
HiCMAE [50]	✓	A+V	81	69.40	68.80
HQC-MAE (VQCL)	✓	A+V	136	**72.90**	**72.85**

Table 7: Average codebook perplexity and normalized utilization across codebook sizes.

Modality	Avg. PPL (Init)	Avg. PPL (End)	U_{Init}	U_{End}
Video ($K_v = 256$)	199.7	251.8	95.5%	99.7%
Video ($K_v = 512$)	312.1	495.1	92.1%	99.5%
Video ($K_v = 1024$)	417.8	956.3	87.1%	99.0%
Video ($K_v = 2048$)	518.7	1775.7	82.0%	98.1%
Audio ($K_a = 160$)	92.3	154.7	89.2%	99.3%
Audio ($K_a = 320$)	127.7	298.8	84.1%	98.8%
Audio ($K_a = 512$)	154.1	457.7	80.8%	98.2%
Audio ($K_a = 1024$)	173.3	805.8	74.4%	96.5%

noise and +11.62%/+5.26% under severe noise, suggesting that incorporating VQ into cross-modal contrastive learning produces more robust representations. HQC-MAE also outperforms the supervised baseline RAVER by +2.49%/+3.84% and +3.28%/+14.23%, respectively, showing that self-supervised pretraining with discrete cross-modal alignment improves robustness under noisy test conditions.

5.5 Cross-Corpus Evaluation

To assess whether introducing VQ into contrastive learning improves generalization, we finetune the HQC-MAE (VQCL) variant on MSP-IMPROV and directly testing on CREMA-D. We keep only the four shared emotion categories between the two corpora: Anger, Happiness, Neutral, and Sadness. The CREMA-D labels are remapped to the same label space as MSP-IMPROV. This setting is more challenging because the training and test data differ in speaker population and recording conditions.

As shown in Table 6, HQC-MAE (VQCL) achieves the best cross-corpus performance. Compared with HiCMAE, HQC-MAE improves performance by +3.50% UAR and +4.05% WAR. It also outperforms the supervised baseline RAVER by +6.78% UAR and +7.30% WAR. These results suggest that the discrete cross-modal representations learned by HQC-MAE transfer better across corpora.

5.6 Codebook Perplexity after Pretraining

To assess how effectively the codebooks use their entries during pretraining, we compute codebook perplexity. For a codebook of size K and a batch of $T = B \times N$ tokens, each token t is assigned to one code via Gumbel-Softmax with hard straight-through estimation, producing a one-hot vector $\mathbf{h}_t \in \{0,1\}^K$. The empirical usage of code k is

$$p_k = \frac{1}{T} \sum_{t=1}^{T} h_{t,k}, \quad k = 1, \ldots, K, \tag{13}$$

and the codebook perplexity is

$$\text{PPL} = \exp\left(- \sum_{k=1}^{K} p_k \log p_k \right). \tag{14}$$

Perplexity ranges from 1 (complete collapse) to K (uniform use of all codes), and can be interpreted as the effective number of active codebook entries. We also report the normalized utilization

$$U = \frac{\ln \text{PPL}}{\ln K}, \tag{15}$$

where $U \in [0, 1]$. We report the average codebook perplexity and normalized utilization at the beginning and end of pretraining for video and audio codebooks with different codebook sizes, as summarized in Table 7. The averages are computed over VQCL-3, VQCL-7, VQCL-11, and FPVQ, where higher perplexity values closer to the codebook size indicate better codebook usage.

For both modalities, the average perplexity increases substantially during pretraining across all codebook sizes, showing that the codebooks remain well utilized rather than collapsing. We observe that normalized utilization decreases as the codebook size increases, indicating a trade-off between representational capacity and codebook usage efficiency. The learned discrete units also exhibit emotion-selective behavior; e.g., one code is selected for 99.9% of happy samples but is rarely selected for sad samples, suggesting that the codebooks capture semantically meaningful discrete units.

6 Conclusions

In this paper, we presented HQC-MAE, a self-supervised learning framework for audio-visual emotion recognition. HQC-MAE is a first attempt to introduce vector quantization at intermediate layers of modality-specific encoders and the cross-modal fusion encoder. We further explore quantized cross-modal contrastive learning to encourage audio and video samples to align under a discrete structure. Due to computational constraints, we use lightweight pretrained encoders. Scaling to larger backbones remains promising but computationally expensive. Future work will extend HQC-MAE to other multimodal tasks, such as audio-visual speaker diarization, where discrete cross-modal alignment is also beneficial.

7 Responsible Innovation Statement

This paper aims to advance self-supervised multimodal representation learning for audio-visual emotion recognition. This technology has many beneficial applications in healthcare and education, but it also carries risks if misused for surveillance and workplace monitoring without informed consent. We encourage responsible implementation and deployment with clear consent, transparency, and fairness considerations.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems* 34 (2021), 24206–24221.

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in neural information processing systems* 33 (2020), 9758–9770.

[3] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*. 609–617.

[4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. 2022. MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*. Springer, 348–367.

[5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*. PMLR, 1298–1312.

[6] Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453* (2019).

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

[8] Luis Bravo, Ciro Rodriguez, Pedro Hidalgo, and Cesar Angulo. 2025. A Systematic Review on Artificial Intelligence-Based Multimodal Dialogue Systems Capable of Emotion Recognition. *Multimodal Technologies and Interaction* 9, 3 (2025), 28.

[9] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*. 205–211.

[10] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. 2017. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing* 8, 1 (January-March 2017), 67–80. doi:10.1109/TAFFC.2016.2515617

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

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

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[14] Hao Cheng, Zhiwei Zhao, Yichao He, Zhenzhen Hu, Jia Li, Meng Wang, and Richang Hong. 2025. VAEmo: Efficient representation learning for visual-audio emotion with knowledge injection. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 5547–5556.

[15] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622* (2018).

[16] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

[18] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

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

[20] L. Goncalves and C. Busso. 2022. Robust Audiovisual Emotion Recognition: Aligning Modalities, Capturing Temporal Information, and Handling Missing Features. *IEEE Transactions on Affective Computing* 13, 4 (October-December 2022), 2156–2170. doi:10.1109/TAFFC.2022.3216993

[21] L. Goncalves, H.-C. Chou, A.N. Salman, C.-C Lee, and C. Busso. 2025. Jointly Learning from Unimodal and Multimodal-Rated Labels in Audio-Visual Emotion Recognition. *IEEE Open Journal of Signal Processing* 6 (January 2025), 165–174. doi:10.1109/OJSP.2025.3530274

[22] L. Goncalves, H.-C. Chou, A. N. Salman, C.-C. Lee, and C. Busso. 2026. Contextual Attention for Robust Audio-Visual Emotion Recognition. *IEEE Open Journal of Signal Processing* early access (2026). doi:10.1109/OJSP.2025.3648710

[23] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso. 2025. Versatile Audio-Visual Learning for Emotion Recognition. *IEEE Transactions on Affective Computing* 16, 1 (January-March 2025), 306–318. doi:10.1109/TAFFC.2024.3433386

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

[25] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 976–980.

[26] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 87–110.

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

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[29] D Hendrycks. 2016. Gaussian Error Linear Units (Gelus). *arXiv preprint arXiv:1606.08415* (2016).

[30] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. 2023. MAViL: Masked audio-video learners. *Advances in neural information processing systems* 36 (2023), 20371–20393.

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

[32] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[33] Simon Jenni, Alexander Black, and John Collomosse. 2023. Audio-visual contrastive learning with temporal self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 7996–8004.

[34] Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. 2024. WavTokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532* (2024).

[35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[36] Brett Koonce. 2021. MobileNetV3. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*. Springer, 125–144.

[37] Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems* 31 (2018).

[38] Milan Lazic, Earl Woodruff, and Jenny Jun. 2025. The Next Generation of Personalized Educational AI: Integrating Emotion and Cognition. In *Social Robots in Education: How to Effectively Introduce Social Robots into Classrooms*. Springer, 63–82.

[39] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before Fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[40] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*. Springer, 280–296.

[41] Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy* 25, 10 (2023), 1440.

[42] Wei-Cheng Lin and Carlos Busso. 2021. Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling. *IEEE Transactions on Affective Computing* 14, 2 (2021), 1215–1227.

[43] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. 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*. 24–32.

[44] Domitilla Magni, Giovanna Del Gaudio, Armando Papa, and Valentina Della Corte. 2024. Digital humanism and artificial intelligence: the role of emotions beyond the human–machine interaction in Society 5.0. *Journal of Management History* 30, 2 (2024), 195–218.

[45] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12475–12486.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision.

In *International conference on machine learning*. PmLR, 8748–8763.

[47] Samir Sadok, Simon Leglaive, and Renaud Séguier. 2025. A vector quantized masked autoencoder for audiovisual speech emotion recognition. *Computer Vision and Image Understanding* 257 (2025), 104362.

[48] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15638–15650.

[49] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. 2023. The effectiveness of MAE pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5484–5494.

[50] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2024. HiCMAE: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion* 108 (2024), 102382.

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

[52] Ioannis Tsiamas, Santiago Pascual, Chunghsin Yeh, and Joan Serrà. 2025. Sequential contrastive audio-visual learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[53] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing* 11, 8 (2017), 1301–1309.

[54] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[55] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14549–14560.

[56] Hanwei Wu and Markus Flierl. 2019. Learning product codebooks using vector-quantized autoencoders for image retrieval. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1–5.

[57] Hanwei Wu and Markus Flierl. 2020. Vector quantization-based regularization for autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6380–6387.

[58] Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, and Jinyu Li. 2024. TS3-Codec: Transformer-based simple streaming single codec. *arXiv preprint arXiv:2411.18803* (2024).

[59] Xuecheng Wu, Heli Sun, Yifan Wang, Jiayu Nie, Jie Zhang, Yabing Wang, Junxiao Xue, and Liang He. 2025. AVF-MAE++: Scaling Affective Video Facial Masked Autoencoders via Efficient Audio-Visual Self-Supervised Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 9142–9153.

[60] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[61] Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. BigCodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377* (2024).

[62] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084* (2021).

[63] Sarthak Yadav, Sergios Theodoridis, and Zheng-Hua Tan. 2025. AudioMAE++: learning better masked audio representations with SwiGLU FFNs. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[64] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715* (2024).

[65] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2007. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *Proceedings of the 9th international conference on Multimodal interfaces*. 126–133.