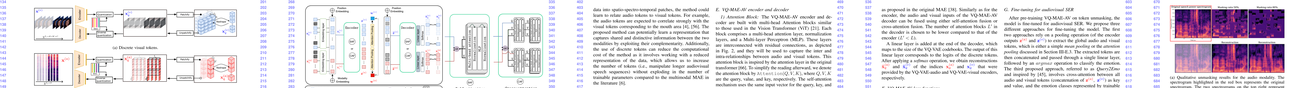


Fig. 1: Discrete audio and visual tokens creation: (i) fully-convolutional VQ-VAEs are trained independently on the audio and visual modalities (see Section III-A); (ii) discrete audio and visual tokens are then produced for multimodal representations provided by the frozen VQ-VAE encoders (see Section III-B).

MAE approach is inspired by masked language modeling [18] and has been successfully applied to image modeling thanks to the development of Vision Transformers (ViT) [21]. The ViT-based MAE approach has also been applied to the time-frequency representation [34], [35]. In the MAE paradigm, the input is divided into non-overlapping patches, each represented by a token embedding. A proportion of tokens are masked, typically 75% for image/audio modeling tasks. The encoder only takes the unmasked tokens as input and fit the masked tokens, and the decoder uses the encoder to reconstruct the masked tokens through a lightweight decoder. To successfully reconstruct the masked tokens, the encoder must generate a semantic representation, which can then be used for downstream tasks. Although the MAE has been recently shown that combining the masking with contrastive learning can improve the audio-only learning [36], [37], [38], [39].

Joint representation learning of audio and visual modalities was presented in [24], [40] for modeling image sequences, called Video-MAE. It uses the same architecture as the vanilla MAE [38] but incorporates a masking process from video ViT (VATT) [41]. Since the joint audio-visual representation learning, particularly in scenes with no motion, the authors propose a cube masking approach along the temporal axis, combined with a high masking ratio of 90%. Other works have extended the MAE to handle multimodal data [6], [29], [35]. Multimae [29] uses multi-task pre-training with a MAE objective from multiple modalities (RGB, depth, and semantic images) and is trained to reconstruct the missing ones. M3AE [35] is a BERT-like architecture that processes image, text, and audio. The main difference in architecture between M3AE and Video-MAE is the modality-specific layers in M3AE. Multiscale Vision Transformers [42] also use a MAE for multimodality, enhancing their fusion by incorporating a cross-attention layer at the outset of each decoder. On the other hand, the M3AE approach uses a single decoder that takes the concatenated tokens from all modalities as input.

Vector-quantized variational autoencoder (VQ-VAE) [43] is a powerful generative model that learns discrete latent vectors. The VQ-VAE encoder takes the input and outputs latent vectors, which are then quantized to the closest embeddings in a learned codebook. The decoder then uses these quantized latent vectors to reconstruct the original input. This is because the VQ-VAE encoder must learn to compress the input into discrete latents that the decoder can accurately reconstruct. The VQ-VAE has been extended to hierarchical variants (VQ-VAE-2) [44] to generate high-fidelity images, and to audio data, where it has been shown that impressive results can be achieved in speech synthesis [45], music generation [46], and general audio generation [47]. The VQ-VAE has also been used for cross-modal generation of images and videos [48], [49], and for multimodal sequences (MIDI files) [50]. Additionally, the VQ-VAE has been used for learning robust speech representations in wav2vec 2.0 [51] and BERT pre-training for image transformers (BEiT) [52].

Masking. We apply masking to the audio and visual sequences of tokens. The VQ-MAE-AV uses masking with a fixed number from the visible tokens to learn a semantic representation of audio and visual tokens and to avoid a bias toward certain downstream tasks. In the original MAE, the masked tokens are randomly drawn with a target ratio of the entire set of tokens. Masking 75% is recommended for images, but this strategy would be to do the same for audio and visual tokens, which would lead to a computationally inefficient and memory demanding pre-training phase, since this strategy would involve randomly drawing a masking proportion $p_m \in [0, 1]$ for each modality, where $p_m = 0.75$ is the visual modality such that $p_a \neq p_v$. To do this, a Dirichlet distribution with three concentration parameters, $\alpha_a = 1.0$, $\alpha_v = 1.0$, and $\alpha_{av} = 4.0$, is used to draw the proportions of each modality, $(p_a, p_v, p_{av}) \sim \mathrm{Dir}(\alpha_a, \alpha_v, \alpha_{av})$. This distribution is uniform overall points in its domain, which are three positive numbers that add up to one. This behavior where most of the tokens are taken from the audio or visual modalities, or even from a combination of both, forces the model to learn a cross-modal representation. Finally, this strategy allows the reconstruction of missing information from a modality given the other two. Moreover, the masking strategy will leverage the less masked modality for the reconstruction of the more masked one. The masking proportions of the two modalities, which can be useful for downstream tasks involving predictions at the sequence level, such as multimodal SER. The global tokens are computed as follows:

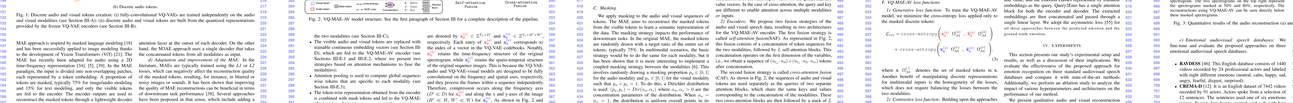
$$\alpha_a^{(i)} \sim \mathrm{Beta}(\alpha_a, \alpha_v)\, p_{av}, \quad \alpha_v^{(i)} = (1 - \alpha_a^{(i)})\, p_{av}, \quad p_a^{(i)} = \alpha_a^{(i)} + p_a, \quad p_v^{(i)} = \alpha_v^{(i)} + p_v \tag{1}$$

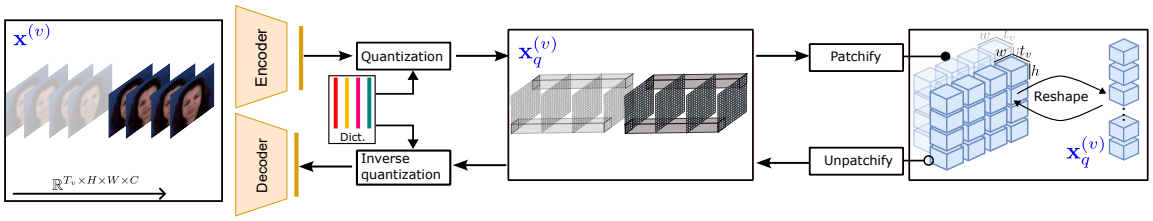
To form positive pairs $(\mathbf{w}^{a_1}, \mathbf{w}^{a_2})$ for both audio and visual modalities, we select overlapping masks between from the same temporal location of two non-overlapping patches to be masked. The VQ-MAE-AV uses the VQ-VAE-audio and VQ-VAE-visual encoders to obtain compressed and quantized representations of the input audio-visual spectrogram pairs $(\mathbf{U}^a, \mathbf{U}^v)$ where $\mathbf{U}^a$ and $\mathbf{U}^v$ correspond to the time and frequency dimensions, and of the embedding dimension D, such that $\mathbf{U}^a \in \mathbb{R}^{T_a \times F_a \times D}$ and $\mathbf{U}^v \in \mathbb{R}^{T_v \times F_v \times D}$. The tokens are then flattened into a sequence of tokens of shape $(T_a \times F_a, D)$ and $(T_v \times F_v, D)$ to be used as input to the VQ-MAE-AV model. The VQ-MAE-AV uses a joint sinusoidal positional embedding between the two modalities (see Section III-C). Let $\mathbf{U}^a \in \mathbb{R}^{N_a \times D}$ and $\mathbf{U}^v \in \mathbb{R}^{N_v \times D}$ denote the sequences of audio and visual tokens, respectively, where $N_a = T_a \times F_a$ and $N_v = T_v \times F_v$. The VQ-MAE-AV model then takes the concatenated tokens from all modalities as input.

Masking strategy. We propose two fusion strategies of the audio and visual spectrogram tokens, resulting in two architectures: (i) VQ-MAE-AV-HiT (hierarchical Transformer) and (ii) VQ-MAE-AV-JoT (joint Transformer). In the original MAE, the masked tokens are randomly drawn with a target ratio of the entire set of tokens. Masking 75% is recommended for images, but this strategy would be to do the same for audio and visual tokens, which would lead to a computationally inefficient and memory demanding pre-training phase, since this strategy would involve randomly drawing a masking proportion $p_m \in [0, 1]$ for each modality, where $p_m = 0.75$ is the visual modality such that $p_a \neq p_v$. To do this, a Dirichlet distribution with three concentration parameters, $\alpha_a = 1.0$, $\alpha_v = 1.0$, and $\alpha_{av} = 4.0$, is used to draw the proportions of each modality, $(p_a, p_v, p_{av}) \sim \mathrm{Dir}(\alpha_a, \alpha_v, \alpha_{av})$. This distribution is uniform overall points in its domain, which are three positive numbers that add up to one. This behavior where most of the tokens are taken from the audio or visual modalities, or even from a combination of both, forces the model to learn a cross-modal representation. Finally, this strategy allows the reconstruction of missing information from a modality given the other two. Moreover, the masking strategy will leverage the less masked modality for the reconstruction of the more masked one. The masking proportions of the two modalities, which can be useful for downstream tasks involving predictions at the sequence level, such as multimodal SER. The global tokens are computed as follows:

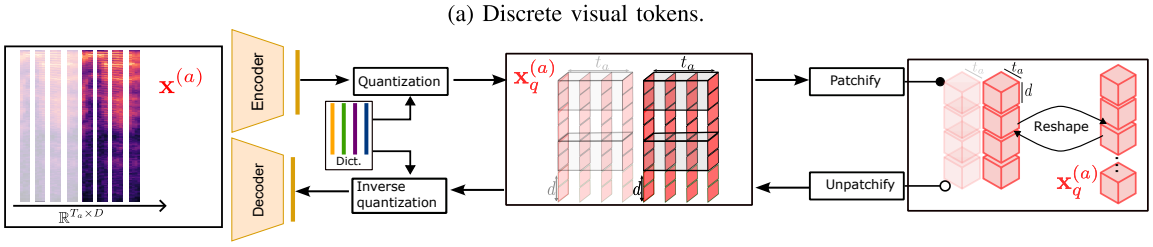
$$\alpha_a^{(i)} \sim \mathrm{Beta}(\alpha_a, \alpha_v)\, p_{av}, \quad \alpha_v^{(i)} = (1 - \alpha_a^{(i)})\, p_{av}, \quad p_a^{(i)} = \alpha_a^{(i)} + p_a, \quad p_v^{(i)} = \alpha_v^{(i)} + p_v \tag{1}$$

To form positive pairs $(\mathbf{w}^{a_1}, \mathbf{w}^{a_2})$ for both audio and visual modalities, we select overlapping masks between from the same temporal location of two non-overlapping patches to be masked. The VQ-MAE-AV uses the VQ-VAE-audio and VQ-VAE-visual encoders to obtain compressed and quantized representations of the input audio-visual spectrogram pairs $(\mathbf{U}^a, \mathbf{U}^v)$ where $\mathbf{U}^a$ and $\mathbf{U}^v$ correspond to the time and frequency dimensions, and of the embedding dimension D, such that $\mathbf{U}^a \in \mathbb{R}^{T_a \times F_a \times D}$ and $\mathbf{U}^v \in \mathbb{R}^{T_v \times F_v \times D}$. The tokens are then flattened into a sequence of tokens of shape $(T_a \times F_a, D)$ and $(T_v \times F_v, D)$ to be used as input to the VQ-MAE-AV model. The VQ-MAE-AV uses a joint sinusoidal positional embedding between the two modalities (see Section III-C). Let $\mathbf{U}^a \in \mathbb{R}^{N_a \times D}$ and $\mathbf{U}^v \in \mathbb{R}^{N_v \times D}$ denote the sequences of audio and visual tokens, respectively, where $N_a = T_a \times F_a$ and $N_v = T_v \times F_v$. The VQ-MAE-AV model then takes the concatenated tokens from all modalities as input.


Masked tokens
Cross-attention
Self-attention
Feed-forward
Concatenation
Linear projection
Classifier
Masked tokens
Masked tokens
Masked tokens



(a) Discrete visual tokens.



(b) Discrete audio tokens.

Fig. 1: Discrete audio and visual tokens creation: (i) fully-convolutional VQ-VAEs are trained independently on the audio and visual modalities (see Section III-A); (ii) discrete audio and visual tokens are built from the quantized representations provided by the frozen VQ-VAE encoders (see Section III-B).

MAE approach is inspired by masked language modeling [18] and has been successfully applied to image modeling thanks to the development of Vision Transformers (ViT) [21]. The MAE has recently been adapted for audio using a 2D time-frequency representation [34], [5], [39]. In the MAE paradigm, the input is divided into non-overlapping patches, each represented by a token embedding. A proportion of tokens are masked, typically 75% for image/audio modeling and 15% for text modeling, and only the visible tokens are fed to the encoder. The encoder outputs are used to reconstruct the masked tokens through a lightweight decoder. To successfully reconstruct the masked tokens, the encoder must generate a semantic representation, which can then be effectively transferred to downstream tasks [38]. It was recently shown that combining the task of reconstructing masked tokens with contrastive learning can improve the representation learned by an MAE [40], [34].

c) Extensions of MAE for sequential and multimodal representation learning: A recent extension of the MAE was presented in [24], [61] for modeling image sequences, called Video-MAE. It uses the same architecture as the vanilla MAE [38] but incorporates a masking process from video ViT (ViViT) [3]. Since videos often contain redundant information, particularly in scenes with no motion, the authors propose a cubic masking approach along the temporal axis, combined with a high masking ratio of 90%. Other works have extended the MAE to handle multimodal data [6], [29], [35]. Multi-MAE [6] encodes a small random subset of visible tokens from multiple modalities (RGB, depth, and semantic images) and is trained to reconstruct the missing ones. M3AE [29] is a unified MAE architecture for two input modalities (image and text). The main difference in architecture between M3AE and MultiMAE lies in the architecture of the decoder. The MulitMAE approach introduces individual decoders for each modality, enhancing their fusion by incorporating a cross-attention layer at the outset of each decoder. On the other hand, the M3AE approach uses a single decoder that takes the concatenated tokens from all modalities as input.

d) Adaptation and improvement of the MAE: In the literature, MAEs are typically trained using the *L1* or *L2* losses, which can negatively affect the reconstruction quality of the masked tokens, resulting, for instance, in blurred or noisy images or sounds. It has been shown that improving the quality of MAE reconstructions can be beneficial in terms of downstream task performance [38]. Several approaches have been proposed in that sense, which include adding a perceptual loss [20] or using discrete representations obtained from VQ generative adversarial networks (VQ-GANs) [23] or variational autoencoders (VQ-VAEs) [65] to train the MAE [43], [57]. These works only considered an unimodal setting, while the present paper proposes a multimodal MAE for audiovisual speech representation learning. Recently, a model called CAV-MAE [35] was proposed combining MAEs and contrastive learning to learn a representation from audiovisual data. However, different from the proposed VQ-MAE-AV model, CAV-MAE processes raw audiovisual data instead of vector-quantized latent representations, it uses simple mean pooling strategies instead of attention pooling, and it is not applied to SER.

III. THE VQ-MAE-AV MODEL

This section presents the VQ-MAE-AV model. The overall approach is illustrated in Fig. 1 and 2 and it can be summarized as follows:

- Fully-convolutional VQ-VAEs are trained independently on the audio and visual modalities (see Section III-A);
- Discrete audio and visual tokens are built from the quantized representations provided by the frozen VQ-VAE encoders (see Section III-B);
- A proportion of the discrete audio and visual tokens is masked out, using a coupled masking strategy between

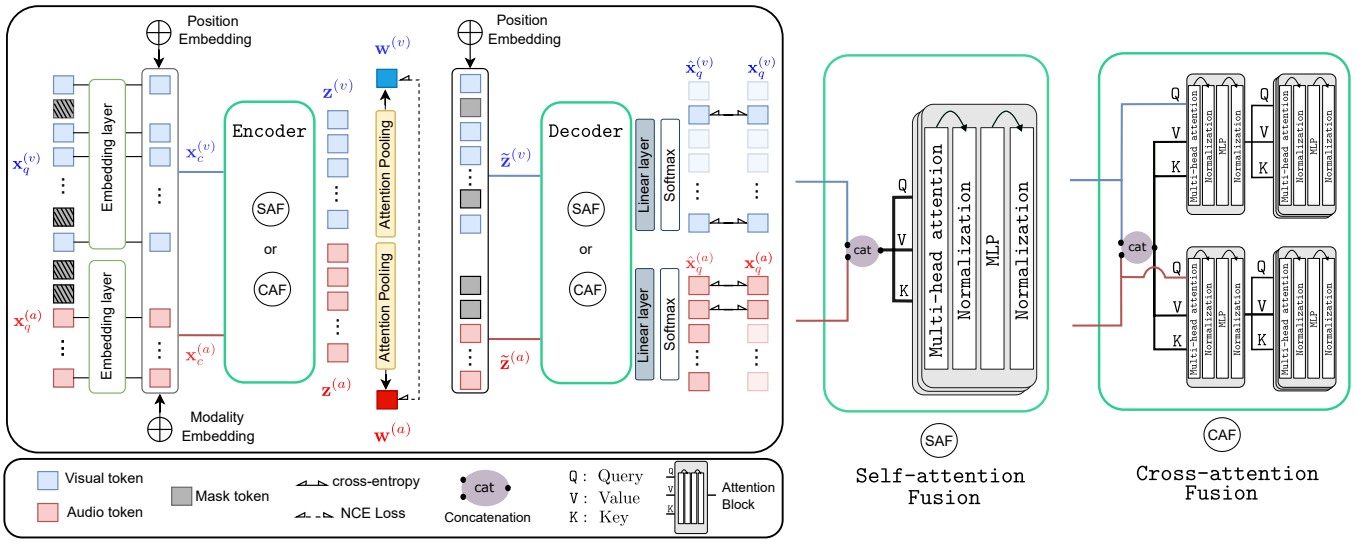


Fig. 2: VQ-MAE-AV model structure. See the first paragraph of Section III for a complete description of the pipeline.

the two modalities (see Section III-C);

- The visible audio and visual tokens are replaced with trainable continuous embedding vectors (see Section III-D), which are fed to the VQ-MAE-AV encoder (see Sections III-E.1 and III-E.2, where we present two strategies based on attention mechanisms to fuse the modalities);
- Attention pooling is used to compute global sequence-wise tokens that are specific to each modality (see Section III-E.3);
- The token-wise representation obtained from the encoder is combined with mask tokens and fed to the VQ-MAE-AV decoder, which tries to reconstruct the original non-masked discrete audio and visual tokens (see Section III-E.4);
- The VQ-MAE-AV model is trained in a self-supervised manner to minimize (i) the cross-entropy loss between the reconstructed and original tokens and (ii) a contrastive loss between the audio and visual global tokens (see Section III-F);
- After self-supervised learning, the VQ-MAE-AV encoder and attention pooling layers are fine-tuned for supervised audiovisual SER (see Section III-G).

This section will present each above-listed aspect of the model in more detail.

A. Vector quantized variational autoencoder

The proposed multimodal self-supervised approach uses the discrete latent representation of two pre-trained and frozen VQ-VAEs [65]. Specifically, as illustrated in Fig. 1, we use the VQ-VAE-audio and VQ-VAE-visual encoders to obtain compressed and quantized representations of the input speech power spectrogram $\mathbf{x}^{(a)} \in \mathbb{R}^{T_a \times D}$ where T_a and D correspond to the time and frequency dimensions, and of the input image sequence $\mathbf{x}^{(v)} \in \mathbb{R}^{T_v \times H \times W \times C}$ where T_v, H, W and C correspond to the time, height, width, and channel dimensions. The audio and visual quantized representations are denoted by $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{T_a \times D'}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{T_v \times H' \times W'}$, respectively. Each entry of $\mathbf{x}_q^{(a)}$ and $\mathbf{x}_q^{(v)}$ corresponds to the index of a vector in the VQ-VAE codebooks. Notably, $\mathbf{x}_q^{(a)}$ retains the time-frequency structure of the original spectrogram, while $\mathbf{x}_q^{(v)}$ retains the spatio-temporal structure of the original sequence images. This is because the VQ-VAE-audio and VQ-VAE-visual models are designed to be fully convolutional on the frequency and spatial axes, respectively, and they process the frames within a sequence independently. Therefore, compression occurs along the frequency axis ($D' \ll D$) for $\mathbf{x}_q^{(a)}$ and along the x and y-axes of the image ($H' \ll H$, $W' \ll W$) for $\mathbf{x}_q^{(v)}$. As shown in Fig. 2 and discussed in the following subsections, the proposed MAE-based self-supervised learning approach operates on these discrete and compressed representations before audiovisual speech reconstruction using the VQ-VAE decoders.

The training procedure of the VQ-VAEs follows the original approach presented in [65]. In particular, the VQ-VAE loss functions involve a reconstruction term between the original and reconstructed data, which corresponds to the mean squared error for the visual modality and to the Itakura-Saito divergence for the audio modality [26]. More details are provided in Section IV-A.2.

B. Discrete audio and visual tokens

As shown in Fig. 1, the audio and visual quantized representations $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{T_a \times D'}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{T_v \times H' \times W'}$ from the output of the VQ-VAE encoders are divided into non-overlapping patches to build discrete tokens $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{(n_{t_a} \cdot t_a) \times (n_d \cdot d)}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{(n_{t_v} \cdot t_v) \times (n_h \cdot h) \times (n_w \cdot w)}$, where $T_a = n_{t_a} \cdot t_a$, $D' = n_d \cdot d$, $T_v = n_{t_v} \cdot t_v$, $H' = n_h \cdot h$, and $W' = n_w \cdot w$. These representations are reshaped to $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{(n_{t_a} \cdot n_d) \times (t_a \cdot d)}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{(n_{t_v} \cdot n_h \cdot n_w) \times (t_v \cdot h \cdot w)}$, which are seen as sequences of $n_{t_a} \cdot n_d$ and $n_{t_v} \cdot n_h \cdot n_w$ tokens of dimension $t_a \cdot d$ and $t_v \cdot h \cdot w$, respectively.

The use of discrete audio and visual tokens in VQ-MAE-AV has several motivations. Firstly, by dividing the audiovisual

data into spatio-spectro-temporal patches, the method could learn to relate audio tokens to visual tokens. For example, the audio tokens are expected to correlate strongly with the visual tokens corresponding to the mouth area [4], [56]. The proposed method can potentially learn a representation that captures shared and distinctive information between the two modalities by exploiting their complementarity. Additionally, the use of discrete tokens can reduce the computational cost of the method as it involves working with a reduced representation of the data, which allows us to increase the number of tokens (i.e., manipulate longer audiovisual speech sequences) without exploding in the number of trainable parameters compared to the multimodal MAE in the literature [6].

C. Masking

We apply masking to the audio and visual sequences of tokens. The MAE aims to reconstruct the masked tokens from the visible tokens to learn a semantic representation of the data. The masking strategy impacts the performance of downstream tasks. In the original MAE, the masked tokens are randomly drawn with a target ratio of the entire set of tokens, typically 75%. In multimodal scenarios, the basic strategy would be to do the same for each modality, but it has been shown that it is more interesting to implement a coupled masking strategy between the modalities [6]. This involves randomly drawing a masking proportion $p_a \in [0,1]$ for the audio modality and $p_v \in [0,1]$ for the visual modality such that $p_a + p_v = 1$. To do this, a Dirichlet distribution is used: $(p_a, p_v) \sim Dir(\alpha_a, \alpha_v)$ where $\alpha_a, \alpha_v > 0$ are the concentration parameters of the distribution. When $\alpha_a = \alpha_v = 1$, the distribution is uniform overall points in its support (the 1-simplex), and $\alpha_a \gg \alpha_b$ results in a sampling behavior where most of the tokens are taken from the audio modality, and vice versa. These parameters can be used to reduce the dominance of one modality over the other. This strategy allows the reconstruction of missing information from one modality by relying on another. Therefore, the model will leverage the less masked modality for the reconstruction of the more masked one. For this paper, we set $\alpha_a = \alpha_v = 1$.

D. Continuous embedding vectors

The discrete tokens correspond to the indices obtained through the quantization step of the pretrained VQ-VAE encoder. Before being input to the VQ-MAE-AV encoder, these discrete tokens are replaced with trainable continuous embedding vectors taken from an audio codebook in $\mathbb{R}^{k_a \times e_a}$ and from a visual codebook in $\mathbb{R}^{k_v \times e_v}$, where $k_{a/v}$ is the number of codes in the codebook and $e_{a/v}$ is the dimension of each code. This is simply achieved by replacing the indices of a discrete token with the corresponding vectors of dimension $e_{a/v}$ in the codebook. After this embedding process, the sequences of discrete tokens $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{(n_{t_a} \cdot n_d) \times (t_a \cdot d)}$ and $\mathbf{x}_q^{(v)}$ in $\mathbb{Z}^{(n_{t_v} \cdot n_h \cdot n_w) \times (t_v \cdot h \cdot w)}$ are transformed into sequences of continuous tokens $\mathbf{x}_c^{(a)} \in \mathbb{R}^{(n_{t_a} \cdot n_d) \times (t_a \cdot d \cdot e_a)}$ and $\mathbf{x}_c^{(v)} \in \mathbb{Z}^{(n_{t_v} \cdot n_h \cdot n_w) \times (t_v \cdot h \cdot w \cdot e_v)}$.

E. VQ-MAE-AV encoder and decoder

1) Attention Block: The VQ-MAE-AV encoder and decoder are built with multi-head Attention blocks similar to those used in the Vision Transformer (ViT) [21]. Each block comprises a multi-head attention layer, normalization layers, and a Multi-layer Perceptron (MLP). These layers are interconnected with residual connections, as depicted in Fig. 2, and they will be used to capture the inter and intra-relationships between audio and visual tokens. This attention block is inspired by the attention layer in the original transformer [66]. To simplify the reading afterward, we denote the attention block by $\texttt{Attention}(Q, V, K)$, where Q, V, K are the query, value, and key, respectively. The self-attention mechanism uses the same input vector for the query, key, and value vectors. In the case of cross-attention, the query and key are different to enable attention across multiple modalities *or* inputs.

2) Encoders: We propose two fusion strategies of the audio and visual speech data, resulting in two architectures for the VQ-MAE-AV encoder. The first fusion strategy is called *self-attention fusion*(SAF). As represented in Fig. 2, this fusion consists of a concatenation of token sequences for the two modalities, followed by L self-attention blocks. This concatenation operates on the first dimension of the variables, i.e., we obtain a sequence of $(n_{t_a} \cdot n_d) + (n_{t_v} \cdot n_h \cdot n_w)$ tokens after concatenation.

The second fusion strategy is called *cross-attention fusion* (CAF). As shown in Fig. 2, the sequences of audio and visual tokens are used separately as the queries of two separate attention blocks, which share the same keys and values corresponding to the concatenation of the modalities. These two cross-attention blocks are then followed by a stack of L self-attention blocks.

For both fusion strategies, the encoder outputs one sequence of tokens for each modality, denoted by $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(v)}$.

3) Global tokens: In addition to the token-wise representations $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(v)}$, we learn two sequence-wise global tokens, denoted by $\mathbf{w}^{(a)} \in \mathbb{R}^{1 \times (t_a \cdot d \cdot e_a)}$ and $\mathbf{w}^{(v)} \in \mathbb{R}^{1 \times (t_v \cdot h \cdot w \cdot e_v)}$, which can be thought of as similar to [CLS] tokens [38]. These global modality-specific tokens are introduced to aggregate the spectro-temporal and spatio-temporal information in the two modalities, which can be useful for downstream tasks involving predictions at the sequence level, such as audiovisual SER. The global tokens are computed as follows:

$$\mathbf{w}^{(a)} = \texttt{Attention}\left(Q_{(a)}, V_{(a)}, K_{(a)}\right) \quad ; \tag{1}$$

$$\mathbf{w}^{(v)} = \texttt{Attention}\left(Q_{(v)}, V_{(v)}, K_{(v)}\right) \quad , \tag{2}$$

where $Q_{(a)} = \mathcal{Z}^{(a)} \in \mathbb{R}^{1 \times (t_a \cdot d \cdot e_a)}$ and $Q_{(v)} = \mathcal{Z}^{(v)} \in \mathbb{R}^{1 \times (t_v \cdot h \cdot w \cdot e_v)}$ represent respectively trainable audio and visual tokens, as proposed in [62], $V_{(a)} = K_{(a)} = \mathbf{z}^{(a)}$, and $V_{(v)} = K_{(v)} = \mathbf{z}^{(v)}$.

4) Decoders: The token-wise representation obtained from the encoder is combined with mask tokens and fed to the VQ-MAE-AV decoder along with additional position embeddings, as denoted by $\tilde{\mathbf{z}}^{(a)}$ and $\tilde{\mathbf{z}}^{(v)}$ and illustrated in Fig. 2. The mask tokens actually correspond to one single trainable vector

as proposed in the original MAE [38]. Similarly as for the encoder, the audio and visual inputs of the VQ-MAE-AV decoder can be fused using either self-attention fusion or cross-attention fusion. The number of attention blocks L' in the decoder is chosen to be lower compared to that of the encoder ($L' < L$).

A linear layer is added at the end of the decoder, which maps to the size of the VQ-VAE codebooks. The output of this linear layer corresponds to the logits of the discrete tokens. After applying a *softmax* operation, we obtain reconstructions $\hat{\mathbf{x}}_q^{(a)}$ and $\hat{\mathbf{x}}_q^{(v)}$ of the indices $\mathbf{x}_q^{(a)}$ and $\mathbf{x}_q^{(v)}$ that were provided by the VQ-VAE-audio and VQ-VAE-visual encoders, respectively.

F. VQ-MAE-AV loss functions

1) Generative loss function: To train the VQ-MAE-AV model, we minimize the cross-entropy loss applied only to the masked discrete tokens:

$$\mathcal{L}_{rec} = \texttt{cross-entropy}\left(\mathbf{x}_q^{(a)} \;\; \Omega_{\mathcal{M}}^{(a)} \;\; , \; \hat{\mathbf{x}}_q^{(a)} \;\; \Omega_{\mathcal{M}}^{(a)}\right) + \texttt{cross-entropy}\left(\mathbf{x}_q^{(v)} \;\; \Omega_{\mathcal{M}}^{(v)} \;\; , \; \hat{\mathbf{x}}_q^{(v)} \;\; \Omega_{\mathcal{M}}^{(v)}\right), \tag{3}$$

where $\mathbf{x} \;\; \Omega_{\mathcal{M}}^{(\cdot)}$ denotes the set of masked tokens in $\mathbf{x}$. Another benefit of manipulating discrete representations for multimodal inputs is the homogeneity of the losses, which does not require balancing the losses between the two modalities.

2) Contrastive loss function: Building upon the approaches presented in [2], [1], the global tokens are learned using noise contrastive estimation. This approach enhances the alignment of audiovisual speech pairs by grouping together embeddings that belong to the same time sequence and separating them from those that do not correspond to the same sequence. This approach involves minimizing the loss function $\mathcal{L}_{NCE}(\mathbf{w}^{(a)}, \mathbf{w}^{(v)})$, which is defined by:

$$\mathcal{L}_{NCE}(\mathbf{u}, \mathbf{v}) = -\log\left(\frac{\exp \;\; \mathbf{u}^\top \mathbf{v}/\tau}{\exp \;\; \mathbf{u}^\top \mathbf{v}/\tau \; + \sum\limits_{(\mathbf{u}', \; \mathbf{v}') \in \mathcal{N}} \exp \;\; \mathbf{u}'^\top \mathbf{v}'/\tau}\right). \tag{4}$$

To form positive pairs $(\mathbf{w}^{(a)}, \; \mathbf{w}^{(v)})$ for both audio and visual modalities, we select corresponding streams from the same temporal location in the video. Conversely, negative pairs $(\mathbf{w}^{'(a)}, \; \mathbf{w}^{'(v)})$ are formed by selecting *non-*corresponding streams drawn from a set $\mathcal{N}$ of different temporal locations for each batch. The sensitivity of the NCE loss in distinguishing between positive and negative pairs is regulated by a temperature parameter $\tau \in \mathbb{R}$.

3) Final loss function: The final loss function corresponds to the sum of the generative loss function in (3) and of the contrastive loss function in (4).

G. Fine-tuning for audiovisual SER

After pre-training VQ-MAE-AV on token unmasking, the model is fine-tuned for audiovisual SER. We propose three different approaches for fine-tuning the model. The first two approaches rely on a pooling operation (of the encoder outputs $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(v)}$) to extract the global audio and visual tokens, which is either a simple *mean pooling* or the *attention pooling* discussed in Section III-E.3. The extracted tokens are then concatenated and passed through a single linear layer, followed by an $argmax$ operation to classify the emotion. The third proposed approach, referred to as *Query2Emo* and inspired by [45], involves cross-attention between all audio and visual tokens (concatenation of $\mathbf{z}^{(a)}$, $\mathbf{z}^{(v)}$) as key and value, and the emotion classes represented by trainable embeddings as the query. Query2Emo has a single attention block for both the encoder and decoder. The extracted embeddings are then concatenated and passed through a single linear layer. We adopt the asymmetric loss [55] for all these approaches between the predicted emotion and the ground-truth emotion.

IV. EXPERIMENTS

This section presents our study's experimental setup and results, as well as a discussion of their implications. We evaluate the effectiveness of the proposed approach for emotion recognition on three standard audiovisual speech databases and compare it with state-of-the-art methods. Additionally, we perform an ablation study to analyze the impact of various hyperparameters and architectures on the performance of our method.

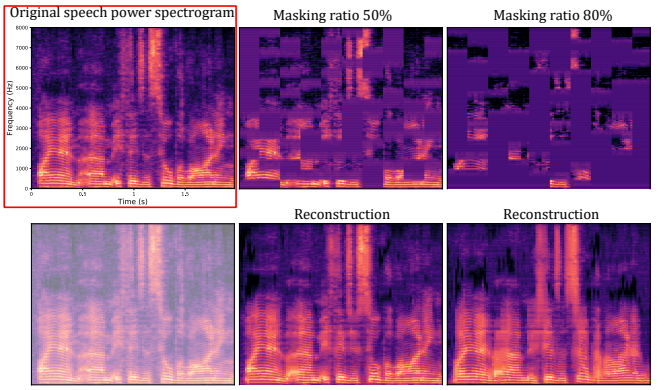
We present qualitative audio and visual reconstruction results in Fig. 3a and Fig. 3b, respectively. The reconstructions are illustrated for two distinct masking ratios: 50% and 80%.

A. Experimental setup

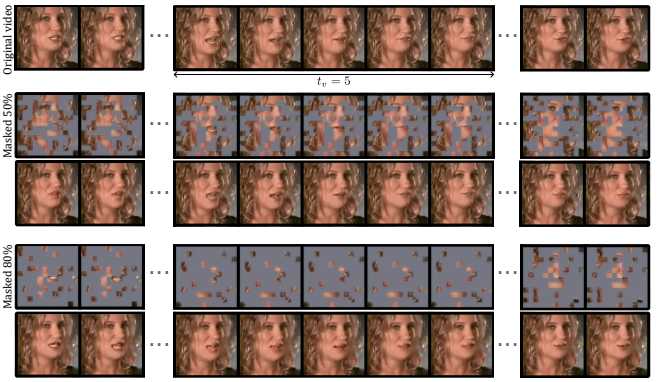
1) Datasets and preprocessing:

a) Dataset for self-supervised training: To pre-train VQ-MAE-AV, we use the VoxCeleb2 dataset [17], which offers a broad range of audiovisual speech data from open-source media, with each video featuring a single speaker. We restricted our dataset use to a subset of around 1000 hours of audiovisual speech, encompassing 2170 different speakers. The test set includes about 100 hours of audiovisual speech data, with 117 different speakers.

b) Data pre-processing: The VQ-VAE-audio and VQ-VAE-visual models are trained on the VoxCeleb2 dataset. The former is trained on short-time Fourier transform (STFT) power spectrograms ($\mathbf{x}^{(a)}$), while the latter is trained on sequences of RGB images ($\mathbf{x}^{(v)}$) captured at 25 fps, cropped and resized to a resolution of 96×96, and aligned using Face-Alignment [10]. To compute the STFT, a Hann window of 64 ms (1024 samples at 16 kHz) and a 68% overlap are used, resulting in a sample rate of 50 Hz, which is twice the sample rate of the visual modality. This leads to sequences of $D = 513$ Fourier coefficients.



(a) Qualitative unmasking results for the audio modality. The spectrogram highlighted in the red box represents the original spectrogram. The two spectrograms on the top right represent the spectrograms masked at 50% and 80%, respectively. The reconstructions using VQ-MAE-AV can be seen directly below these masked spectrograms.



(b) Qualitative unmasking results for the visual modality. The first sequence shows the original video, followed by the next two sequences representing the masked video with a ratio of 50% and its reconstruction using VQ-MAE-AV. The last two sequences represent the masked video with a ratio of 80% and its reconstruction using VQ-MAE-AV.

Fig. 3: Quantitative results of the audio reconstruction (a) and visual reconstruction (b) using the VQ-MAE-AV model.

c) Emotional audiovisual speech databases: We fine-tune and evaluate the proposed approaches on three emotional audiovisual speech databases.

- **RAVDESS** [46]: This English database consists of 1440 videos recorded by 24 professional actors and labeled with eight different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised).
- **CREMA-D** [12]: It is an English dataset of 7442 videos recorded by 91 actors. Actors spoke from a selection of 12 sentences. The sentences used one of six emotions (anger, disgust, fear, happy, neutral, and sad) with four different intensities (low, medium, high, and unspecified).
- **eNTERFACE05** [50]: The audiovisual acted dataset comprises recordings of six distinct emotions: anger, disgust, fear, joy, sadness, and surprise. These emotional expressions were elicited from a diverse group of 43 individuals representing 14 nationalities. The dataset encompasses a total of 1290 video samples. Each participant was instructed to listen to six short stories that elicited a specific emotion.

We selected these databases as they are commonly used for the emotion recognition task, and their raw data is accessible. To ensure a fair comparison with previous works, we performed *6*-fold, *10*-fold and *5*-fold cross-validation for the RAVDESS, CREAM-D and eNTERFACE05 datasets, respectively. We carefully partitioned the datasets to ensure speaker identity separation between fine-tuning and evaluation phases.

2) Model architecture:

a) VQ-VAE architectures: The VQ-VAE-audio (respectively VQ-VAE-visual) architecture is symmetrical concerning the encoder and the decoder, with three 1D (respectively 2D) convolution for the encoder or transposed convolution for the decoder layers and a residual convolution layer. The VQ-VAE models process each frame independently with no time dependency. For each speech power spectrogram frame of size $D = 513$, the VQ-VAE-audio encoder compresses it into a discrete latent vector (a column of $\mathbf{x}_q^{(a)}$) of size $D' = 64$. For each image frame of size $(H = 96, W = 96, C = 3)$, the VQ-VAE-visual encoder compresses it into a discrete latent representation of size $(H' = 24, W' = 24)$. The VQ-VAE-audio and the VQ-VAE-visual codebooks contain, respectively, $k_a = 256$ and $k_v = 512$ codes of dimension $e_a = 8$ and $e_v = 4$. Such a low dimension is chosen to increase the use of the different codes in the codebook [69].

b) VQ-MAE-AV architectures: The VQ-MAE-AV model uses $L = 12$ attention blocks in the encoder and $L' = 4$ in the decoder. Each self-attention layer of a block is divided into *four* heads. By default, the parameters of the discrete audio and visual tokens (d, h, and w) are set to 4. We set $t_a = 10$ and $t_v = 5$ because the sampling rate of $\mathbf{x}^{(a)}$ is twice the sampling rate of $\mathbf{x}^{(v)}$. In the ablation study (Section IV-C), we will explore all possible combinations of the VQ-MAE-AV encoder and decoder, according to the two fusion strategies (*self-attention fusion* and *cross-attention fusion*). We will also evaluate the impact of the pooling strategy and contrastive learning.

3) Training and fine-tuning details:

a) Self-supervised training details: The VQ-MAE-AV is trained using the AdamW optimizer [47] with a cosine scheduler to adjust the learning rate, with a 100-epoch warm-up period. The parameters of the optimizer, similar to [38], are $\beta_2 = 0.9$, $\beta_2 = 0.95$, and $\texttt{weight_decay} = 0.05$. The base learning rate follows the linear scaling rule [36] $lr = (\texttt{base_lr} = 1e-3) \times (\texttt{batchsize} = 128)/256$. We distributed the pre-training of VQ-MAE-AV on 4 NVIDIA HGX A100. The training lasted for 160 epochs, and each epoch took approximately 15 minutes.

b) Fine-tuning details: For the fine-tuning process, we also use the AdamW optimizer [47] with a cosine scheduler to adjust the learning rate and with a 40-epoch warm-up period. The parameters of the optimizer are the same as those

TABLE I: Accuracy (%) and F1-score (%) results of audiovisual SER. VQ-MAE-AV is pre-trained with both the generative (3) and contrastive (4) loss functions using the Query2Emo pooling strategy. In the modality column, A and V stand for audio and visual, respectively. The best scores are in bold and second best scores are underlined.

RAVDESS				CREMA-D				eNTERFACE05			
Method	Modality	Acc	F1	Method	Modality	Acc	F1	Method	Modality	Acc	F1
AV-LSTM [31]	A+V	65.80	-	AV-LSTM [31]	A+V	72.90	-	Graph-Tran [72]	V	54.62	-
MuLT [64]	A+V	76.60	77.30	MATER [30]	A+V	67.20	-	AugAV [48]	A+V	<u>71.13</u>	-
AVT [16]	A+V	79.20	78.20	RAVER [33]	A+V	77.30	-	3D CNN [37]	A	**72.33**	-
MDVAE [56]	A+V	79.30	80.70	MulT Large [63]	A+V	70.22	-	SIAMESE [25]	A	40.00	-
Emotion2vec [49]	A	<u>82.43</u>	-	Wav2Vec2.0 [7]	A	72.41	-	MAE-DFER [59]	V	61.64	-
SVFAP-B [60]	V	75.15	-	MAE-DFER [59]	V	<u>77.38</u>	-	SVFAP-B [60]	V	60.54	-
VQ-MAE-AV (ours)	A	73.20	72.80	VQ-MAE-AV (ours)	A	72.10	71.60	VQ-MAE-AV (ours)	A	65.20	65.70
VQ-MAE-AV (ours)	V	74.10	73.90	VQ-MAE-AV (ours)	V	76.50	76.70	VQ-MAE-AV (ours)	V	60.70	60.50
VQ-MAE-AV (ours)	A+V	**84.80**	**84.50**	VQ-MAE-AV (ours)	A+V	**80.40**	**80.00**	VQ-MAE-AV (ours)	A+V	69.70	69.20

TABLE II: Performance of VQ-MAE-AV using the *self-attention fusion* strategy for both encoder and decoder, and using the attention pooling strategy for the fine-tuning. 'Pre-train' refers to the training of the VQ-MAE-AV for the unmasking task on the VoxCeleb2 database. 'Freeze' refers to the freezing of the VQ-MAE-AV encoder.

Method	Pre-train	Freeze	Accuracy (%)
VQ-MAE-AV	✗	✗	29.6
VQ-MAE-AV	✓	✓	70.5
VQ-MAE-AV	✓	✗	81.5

TABLE III: Performance of VQ-MAE-AV using the *self-attention fusion* strategy for both encoder and decoder, and using the Query2Emo pooling strategy for the fine-tuning. 'Generative' corresponds to the loss function in Eq. 3 and 'Contrastive' corresponds to the loss function in Eq. 4.

Method	Contrastive	Generative	Accuracy (%)
VQ-MAE-AV	✓	✗	75.2
VQ-MAE-AV	✗	✓	84.3
VQ-MAE-AV	✓	✓	84.8

TABLE IV: Performance of VQ-MAE-AV without contrastive learning for different encoder and decoder architectures, using the attention pooling strategy for the fine-tuning. *SAF* stands for *self-attention fusion* and *CAF* stands for *cross-attention fusion*.

Method	Param. (M)	Encoder	Decoder	Acc. (%)
VQ-MAE-AV	13.5	*SAF*	*SAF*	81.5
VQ-MAE-AV	25.0	*CAF*	*SAF*	82.8
VQ-MAE-AV	18.4	*SAF*	*CAF*	81.8
VQ-MAE-AV	30.0	*CAF*	*CAF*	83.0

TABLE V: Performance of VQ-MAE-AV using the *self-attention fusion* strategy for both encoder and decoder, without contrastive learning, and using different pooling strategies for the fine-tuning.

Pooling strategy	Accuracy (%)	F1-score (%)
Mean Pooling	78.1	78.4
Attention Pooling	81.5	80.1
Query2Emo	84.3	84.8

used for the pre-training. The base learning rate is `1e-4`.

B. Audiovisual emotion recognition

Table I compares the emotion recognition performance (accuracy and F1-score metrics) of the proposed VQ-MAE-AV model (using the *cross-attention fusion* strategy for both the encoder and decoder) with the performance of several state-of-the-art methods. These methods are structured in the following manner: supervised approaches are listed in the first block of the table, followed by unsupervised or self-supervised approaches in the second block.

The VQ-MAE-AV model with Query2Emo outperforms the most recent methods overall. VQ-MAE-AV achieves 5.60%, 5.50%, and 2.37% better accuracy than the AVT [16], MDVAE [56], and emotion2vec [49] methods for the RAVDESS dataset, respectively. Regarding the CREMA-D dataset, VQ-MAE-AV achieves 7.99%, 3.10%, and 3.02% better accuracy than the Wav2Vec2.0 [7], RAVER [33], and MAE-DFER [59]. On the eNTERFACE05 dataset, our model outperforms self-supervised approaches, showcasing a significant gain of 8.06% compared to MAE-DFER, and maintains competitiveness with supervised approaches. Overall, the experimental results demonstrate the effectiveness of the proposed audiovisual self-supervised representation learning technique for SER.

Table I also compares VQ-MAE-AV with its unimodal versions VQ-MAE-V and VQ-MAE-A. The VQ-MAE-AV model exhibits improved performance compared to VQ-MAE-V and VQ-MAE-A, with gains of 3.90% and 8.30% in accuracy on the CREMA-D dataset, respectively. This shows the effectiveness of incorporating audiovisual information in self-supervised learning for improved audiovisual SER performance.

C. Ablation study and model properties

We conduct a series of experiments to evaluate the impact of various hyperparameters and model designs on the emotion

recognition performance of the proposed VQ-MAE-AV model. In the following paragraphs, we will present and discuss the findings of these experiments. All the ablation study is done on RAVDESS.

1) Impact of pre-training and fine-tuning: Table II shows the significance of pre-training and fine-tuning the VQ-MAE-AV model for audiovisual SER. Pre-training the model for the unmasking task on the VoxCeleb2 database substantially improves the emotion recognition performance, with the accuracy rising from 29.6% to 81.5%. Fine-tuning the encoder is also essential, as keeping it frozen leads to a 11% drop in accuracy.

2) Impact of the contrastive learning: The influence of contrastive learning on emotion recognition is illustrated in Table III. When training VQ-MAE-AV exclusively with the contrastive loss, the achieved accuracy is 75.2%, which is 9.1% lower than when using only the generative loss. Notably, the combination of both losses, contrastive and generative, leads to improved emotion recognition, resulting in a 0.5% accuracy boost compared to using the generative loss alone. This underscores the importance of integrating both contrastive and generative learning strategies.

3) Impact of the encoder/decoder architecture: Table IV shows the performance of the four different configurations of the VQ-MAE-AV model described in Section III-E, including all possible combinations of the fusion strategies for the VQ-MAE-AV encoder and decoder:

- *Self-attention fusion* for both the encoder and decoder (SAF-SAF);
- *Self-attention fusion* for the encoder and *cross-attention fusion* for the decoder (SAF-CAF);
- *Cross-attention fusion* for the encoder and *Self-attention fusion* for the decoder (CAF-SAF);
- and *Cross-attention fusion* for both the encoder and decoder (CAF-CAF).

Among these configurations, CAF-CAF achieves the highest accuracy with a 1.5% improvement over SAF-SAF, followed by CAF-SAF with a 1.3% improvement over SAF-SAF, and then SAF-CAF with only a 0.3% improvement. The *cross-attention fusion* encoder architecture achieves the best performance in emotion recognition, as shown by the CAF-CAF and CAF-SAF configurations. However, there exists a trade-off between performance and the number of model parameters. Notably, CAF-CAF involves slightly more than twice as many parameters as SAF-SAF.

4) Impact of the pooling strategy: Table V presents the impact of various pooling strategies when fine-tuning the model for audiovisual SER. The results reveal that attention-based pooling techniques, such as attention pooling and Query2Emo, outperform mean pooling. Among the two attention-based methods, Query2Emo outperforms attention pooling, with an accuracy gain of 2.8%.

5) Impact of the audio and visual discrete token size: Fig. 4 shows the impact of the dimensions of the discrete tokens for the visual (h and w) and audio (d) modalities on the audiovisual SER performance. The values of h and w represent the visual token size on the horizontal and vertical axes, respectively, while d represents the token size on the frequency axis of the audio modality. Our study reveals that the performance of emotion recognition is impacted by both (h, w) and d values. Therefore, it is important to select these parameters carefully. Based on our experiments, we recommend fixing them to $(h = w = 4,\ d = 4)$.

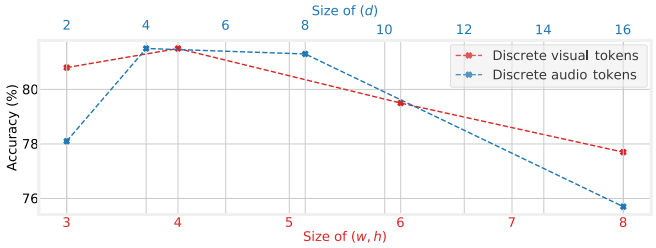


Fig. 4: Impact of the discrete audio and visual token size on emotion recognition.

6) Other ablation studies: We conducted additional ablation studies to investigate the impacts of the encoder depth and of the masking strategy, which we summarize as follows: Increasing the depth of the encoder improves the performance to a certain extent, after 16 blocks the performance starts to decrease. Additionally, comparing masking techniques reveals that the Dirichlet distribution-based masking outperforms fixed ratios for the two modalities by 0.8%, 0.4%, and 1.9% accuracy across different ratios (70%, 80%, and 90%).

V. Conclusion

Masked autoencoder modeling is a versatile self-supervised learning approach that can be adapted to various types of data. This paper introduced the VQ-MAE-AV model for learning representations of audiovisual speech data, which could be extended to other multimodal sequential data. VQ-MAE-AV took as input a discrete audio representation and a discrete visual representation obtained via two separate VQ-VAEs. These reduced representations were then divided into multiple discrete tokens, with spatio-temporal tokens for the visual modality and spectro-temporal tokens for the audio modality. Pre-trained on the VoxCeleb2 dataset and fine-tuned on standard expressive audiovisual speech datasets, the experiments showed that the VQ-MAE-AV model effectively combined the audio and visual modalities for audiovisual SER, outperforming several state-of-the-art methods across multiple datasets.

The experimental validation of the proposed approach was conducted using three well-established datasets in the field of audiovisual SER. It could be interesting to further extend this validation using more recent in-the-wild datasets such as Aff-wild2 [42]. For future work, we plan to investigate other masking strategies and increase the resolution of the images to further improve emotion recognition performance. We could also explore the use of the representation learned by the proposed VQ-MAE-AV model for other downstream tasks in audiovisual speech processing, such as speaker recognition.

REFERENCES

[1] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:24206–24221, 2021.

[2] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:25–37, 2020.

[3] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. ViViT: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021.

[4] M. Arnela, R. Blandin, S. Dabbaghchian, O. Guasch, F. Alías, X. Pelorson, A. Van Hirtum, and O. Engwall. Influence of lips on the production of vowels based on finite element simulations and experiments. *The Journal of the Acoustical Society of America*, 139(5):2852–2859, 2016.

[5] A. Baade, P. Peng, and D. Harwath. MAE-AST: Masked Autoencoding Audio Spectrogram Transformer. In *International Speech Communication Association (Interspeech)*, pages 2438–2442, 2022.

[6] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12449–12460, 2020.

[8] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 19, 2006.

[10] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[12] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[13] L.-W. Chen and A. Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[15] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[16] K. Chumachenko, A. Iosifidis, and M. Gabbouj. Self-attention fusion for audiovisual emotion recognition with incomplete data. In *International Conference on Pattern Recognition (ICPR)*, pages 2822–2828. IEEE, 2022.

[17] J. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *International Speech Communication Association (Interspeech)*, 2018.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[19] A. Dib, J. Ahn, C. Thebault, P.-H. Gosselin, and L. Chevallier. S2f2: Self-supervised high fidelity face reconstruction from monocular image. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

[20] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, and N. Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2020.

[22] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[23] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 12873–12883, 2021.

[24] C. Feichtenhofer, Y. Li, K. He, et al. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:35946–35958, 2022.

[25] K. Feng and T. Chaspari. A siamese neural network with modified distance loss for transfer learning in speech emotion recognition. *arXiv preprint arXiv:2006.03001*, 2020.

[26] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[27] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10478–10487, 2020.

[28] R. Gao and K. Grauman. Co-separating sounds of visual objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3879–3888, 2019.

[29] X. Geng, H. Liu, L. Lee, D. Schuurams, S. Levine, and P. Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

[30] E. Ghaleb, J. Niehues, and S. Asteriadis. Multimodal attention-mechanism for temporal emotion recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 251–255. IEEE, 2020.

[31] E. Ghaleb, M. Popa, and S. Asteriadis. Multimodal and temporal perception of audio-visual cues for emotion recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 552–558. IEEE, 2019.

[32] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[33] L. Goncalves and C. Busso. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170, 2022.

[34] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass. Ssast: Self-supervised audio spectrogram transformer. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022.

[35] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass. Contrastive audio-visual masked autoencoder. In *International Conference on Learning Representations (ICLR)*, 2022.

[36] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch SGD: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[37] N. Hajarolasvadi and H. Demirel. 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, 21(5):479, 2019.

[38] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.

[39] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:28708–28720, 2022.

[40] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13, 2023.

[41] M. Jegorova, S. Petridis, and M. Pantic. SS-VAERR: Self-supervised apparent emotional reaction recognition from video. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

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

[43] T. Li, H. Chang, S. Mishra, H. Zhang, D. Katabi, and D. Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2142–2152, 2023.

[44] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller. Audio self-supervised learning:

A survey. *Patterns*, 3(12):100616, 2022.
[45] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.
[46] S. R. Livingstone and F. 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), 2018.
[47] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2017.
[48] F. Ma, Y. Li, S. Ni, S.-L. Huang, and L. Zhang. Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN. *Applied Sciences*, 12(1):527, 2022.
[49] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.
[50] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACE'05 audio-visual emotion database. In *22nd international conference on data engineering workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.
[51] A. Mehrabian. *Nonverbal communication*. Routledge, 2017.
[52] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision (ECCV)*, pages 69–84. Springer, 2016.
[53] L. Pepino, P. Riera, and L. Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *International Speech Communication Association (Interspeech)*, pages 3400–3404, 2021.
[54] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.
[55] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor. Asymmetric loss for multi-label classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91. IEEE Computer Society, 2021.
[56] S. Sadok, S. Leglaive, L. Girin, X. Alameda-Pineda, and R. Séguier. A multimodal dynamical variational autoencoder for audiovisual speech representation learning. *Neural Networks*, page 106120, 2024.
[57] S. Sadok, S. Leglaive, and R. Séguier. A vector quantized masked autoencoder for speech emotion recognition. In *IEEE ICASSP 2023 Workshop on Self-Supervision in Audio, Speech and Beyond (SASB)*, 2023.
[58] L. Schoneveld, A. Othmani, and H. Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146:1–7, 2021.
[59] L. Sun, Z. Lian, B. Liu, and J. Tao. MAE-DFER: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *International Conference on Multimedia (ACM)*, page 6110–6121, 2023.
[60] L. Sun, Z. Lian, K. Wang, Y. He, M. Xu, H. Sun, B. Liu, and J. Tao. Svfap: Self-supervised video facial affect perceiver. *arXiv preprint arXiv:2401.00416*, 2023.
[61] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:10078–10093, 2022.
[62] H. Touvron, M. Cord, A. El-Nouby, P. Bojanowski, A. Joulin, G. Synnaeve, and H. Jégou. Augmenting convolutional networks with attention-based aggregation. *arXiv preprint arXiv:2112.13692*, 2021.
[63] M. Tran and M. Soleymani. A pre-trained audio-visual transformer for emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4698–4702. IEEE, 2022.
[64] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.
[65] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
[67] Y. Wang, A. Boumadane, and A. Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*, 2021.
[68] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022.
[69] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu. Vector-quantized image modeling with improved vqgan. In *International Conference on Learning Representations (ICLR)*, 2021.
[70] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022.
[71] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba. The sound of motions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1735–1744, 2019.
[72] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam. Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition. *IEEE Transactions on Affective Computing*, 2022.