

VQTalker: Towards Multilingual Talking Avatars Through Facial Motion Tokenization

Tao Liu^{1*}, Ziyang Ma¹, Qi Chen¹, Feilong Chen², Shuai Fan², Xie Chen¹, Kai Yu^{1†}

¹X-LANCE Lab, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University

²AI-Speech Ltd

Abstract

We present VQTalker, a Vector Quantization-based framework for multilingual talking head generation that addresses the challenges of lip synchronization and natural motion across diverse languages. Our approach is grounded in the phonetic principle that human speech comprises a finite set of distinct sound units (phonemes) and corresponding visual articulations (visemes), which often share commonalities across languages. We introduce a facial motion tokenizer based on Group Residual Finite Scalar Quantization (GRFSQ), which creates a discretized representation of facial features. This method enables comprehensive capture of facial movements while improving generalization to multiple languages, even with limited training data. Building on this quantized representation, we implement a coarse-to-fine motion generation process that progressively refines facial animations. Extensive experiments demonstrate that VQTalker achieves state-of-the-art performance in both video-driven and speech-driven scenarios, particularly in multilingual settings. Notably, our method achieves high-quality results at a resolution of 512×512 pixels while maintaining a lower bitrate of approximately 11 kbps. Our work opens new possibilities for cross-lingual talking face generation.

Introduction

Audio-driven talking head generation aims to create realistic facial animation synchronized with input audio and has various applications in film dubbing and animation production. The core challenge lies in precise lip synchronization, which is crucial due to the McGurk effect (McGurk and MacDonald 1976). This effect demonstrates the intricate interaction between auditory and visual cues in human speech perception. Although current audio-driven talking head generation methods have made significant progress, existing algorithms (Zhang et al. 2023a; Ma et al. 2023; Liu et al. 2024) still face synchronization issues, such as misalignment between audio and visual cues or the absence of specific lip shapes. These problems are particularly pronounced when dealing with languages outside the Indo-European family.

This limitation primarily stems from two factors. Commonly used training datasets like VoxCeleb (Nagrani,

Chung, and Zisserman 2017), CelebV-HQ (Zhu et al. 2022), MEAD (Wang et al. 2020), and HDTF (Zhang et al. 2021) are overwhelmingly dominated by Indo-European languages, like English, German, and French. As a result, models trained on these datasets perform well within the language family but often fail to accurately capture the lip movements and facial expressions associated with other linguistic groups. Secondly, the reliance on continuous representations in existing methods further compounds the problem. Those approaches typically operate in continuous spaces, such as the continuous spectral representations of audio inputs and the continuous latent embeddings of facial motion outputs. However, different language families possess a finite number of distinct phonetic structures and visual articulations. Continuous representations allow for infinite variations, potentially increasing training difficulty and leading to overfitting specific languages.

While some approaches (Yang et al. 2020; Huang et al. 2021; Song et al. 2022; Sung-Bin et al. 2024) focus on increasing training data volume for other languages, this strategy is resource-intensive and particularly challenging for minority languages. Moreover, simply increasing data volume may not fully address the underlying need for models that can generalize across diverse linguistic systems.

In this paper, we propose **VQTalker**. Our approach leverages the fundamental linguistic concept of phonemes and visemes to create a novel system for talking head generation. By mapping discretized speech units to discretized facial motion tokens, we develop a method that captures the essential elements of speech-driven facial animation across diverse languages. Each speech unit or combination in our framework is conceptually linked to a phoneme, and each facial motion token is also conceptually tied to a viseme. This discrete representation allows us to model the universal patterns of speech-to-facial movement correlations efficiently. Figure 1 provides a depiction of the core process by which VQTalker converts audio tokens into facial motion tokens, bringing a portrait to life with discrete mouth shapes.

Specifically, we design a facial motion vector tokenizer to model comprehensive facial movements, encompassing head pose, gaze direction, eye blinks, and most critically, the nuanced dynamics of lip shapes. These motions, particularly the intricate variations in lip articulation, form a finite set of possible states. However, directly modeling facial motion at

*Completed during the author’s internship.

†The Corresponding author.

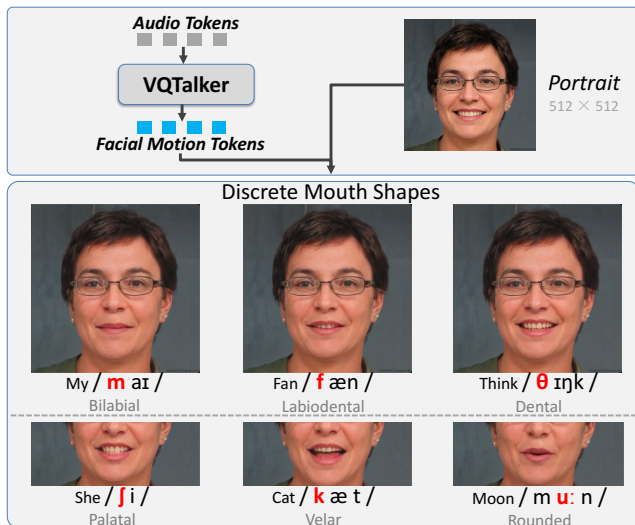


Figure 1: VQTalker converts audio tokens to facial motion tokens, animating a source portrait with diverse discrete mouth shapes. Red text indicates phonetic sounds, with corresponding viseme categories in grey. Synthetic results can be viewed at <https://x-lance.github.io/VQTalker>.

high resolution presents challenges, primarily due to the limited capacity of a single codebook to capture complex facial dynamics. To address this, our implementation employs a multi-strategy quantization method based on Group Residual Finite Scalar Quantization (GRFSQ). This combined approach leverages the strengths of different strategies: group quantization reduces codebook size, residual quantization allows for modeling actions at different granularities in a coarse-to-fine manner, and Finite Scalar Quantization (FSQ) is codebook-free and improves codebook utilization. Finally, this enables us to model 512×512 resolution motion at a low bitrate of approximately 11 kbps, which is only about 70% of the current lowest bitrate of 16 kbps for continuous representation methods (Wang et al. 2022; Liu et al. 2024).

Building on this structure, we introduce an interleaving pattern for codebook generation that incorporates both autoregressive and non-autoregressive elements. Within each motion granularity, we employ a non-autoregressive approach to maintain bidirectional attention, fully leveraging temporal correlations. Between different granularities, we utilize an autoregressive strategy, resulting in a coarse-to-fine motion generation process. This approach strikes a trade-off between generation speed and complexity, ensuring both efficiency and quality in the produced animations.

The main contributions of our work are:

1. A facial motion tokenizer that leverages the finite nature of speech articulation to create a discretized representation, enhancing the capture of subtle lip movements and improving cross-lingual generalization.
2. A Group Residual Finite Scalar Quantization (GRFSQ) method that combines group, residual, and finite scalar quantization strategies to achieve efficient high-

resolution facial motion modeling at low bitrates.

3. A coarse-to-fine motion generation process combining non-autoregressive temporal and autoregressive granular approaches, enhancing both consistency and efficiency.

Related Works

Latent Facial Representation. Recent studies (He et al. 2024; Xu et al. 2024; Drobyshev et al. 2024) have explored more refined representations through self-supervised learning. While this approach models facial features more precisely than predefined structured features like landmarks, it often implicitly encodes target identity information, leading to identity leakage issues. Although various strategies (Drobyshev et al. 2022; Liu et al. 2024) have been proposed to address this issue, they introduce additional model complexity and computational overhead. In this paper, we propose exploring **discrete coding** in Latent Facial Representation. Discrete coding inherently creates an information bottleneck by quantizing the continuous feature space into a finite discrete codebook. This approach enables the model to learn more compact representations and mitigates information leakage without additional modules or losses.

Facial Tokenization. Existing approaches (Xing et al. 2023a; Tan, Ji, and Pan 2024; Yu et al. 2024) typically rely on VQ-VAE (Esser, Rombach, and Ommer 2021), **patch-based** tokenization, which limits to local regions or specific patterns (Xing et al. 2023b; Tan et al. 2024b) and fails to capture global semantics, constraining their application to lower resolutions. Scaling these methods to higher resolutions often requires complex networks with billions of parameters (Tian et al. 2024; Chen et al. 2024). We argue that the drawback lies in coupling texture and motion information. Due to the texture part containing detailed information, the network’s reconstruction of these details leads to an increase in network parameters, making it difficult to scale to higher resolutions or requiring larger networks. In contrast, our method decouples these elements by independently quantizing motion and using a separate rendering network to recover the detailed parts. This approach allows us to transcend local patch semantics to **global motion semantics**, enabling holistic modeling of facial dynamics without the need for excessively large networks.

Vector Quantization (VQ) maps continuous input vectors to a finite set of discrete vectors for compressing data or improving the network efficiency. **Group VQ (GVQ)** (Baeviski, Schneider, and Auli 2020; Chen, Watanabe, and Rudnicky 2023) utilizes multiple groups to model different patterns, alleviating mode collapse in single codebooks. **Residual VQ (RVQ)** (Zeghidour et al. 2021; Copet et al. 2024) quantizes vectors iteratively, encoding the residual error, which is particularly suitable for features with different granularity representations. For instance, in audio processing, shallow semantic features and deep acoustic features can be modeled at different granularities in various residual layers. In terms of codebook learning, compared with original codebook optimization, **Finite Scalar Quantization (FSQ)** (Fabian et al. 2024) optimizes the table lookup process and incorporates the Straight-Through

Estimator (STE) technique into the network. It projects the continuous representation of neural networks onto a few dimensions, then quantizes each dimension to a finite set of values. This approach achieves good performance and better codebook utilization without complex auxiliary losses. Our method will combine the strengths of those methods.

VQTalker Framework

Our proposed VQTalker framework addresses the challenges of multilingual talking head generation by transforming speech tokens into visual motion tokens. As illustrated in Figure 2, the framework consists of two primary components: (a) a quantized facial motion tokenizer, and (b) a coarse-to-fine motion generation process.

Quantized Facial Motion Tokenizer

The core of our approach is a quantized facial motion tokenizer designed specifically for talking head tasks. This tokenizer, depicted in Figure 2 (a), leverages self-supervised learning and a multi-strategy quantization method to create a discretized representation of facial motion, encompassing lip movements, blinking, head pose, and gaze direction.

The tokenizer follows an Encoder-Quantizer-Decoder architecture. The Image Encoder extracts three types of features from both source and target images: fine-grained facial features for high-frequency detail, global facial features f_g serving as a facial prompt, and motion features f_s and f_t from source and target images respectively. The Quantizer converts the motion difference $f_{s \rightarrow t} = f_t - f_s$ into discrete tokens using Group Residual Finite Scalar Quantization (GRFSQ), acting as a strong information bottleneck.

The Image Renderer (Decoder) reconstructs facial features from the quantized representation, combining them with the global face feature f_g to guide the final rendering. To recover fine details, each decoder layer is connected to its corresponding encoder layer through skip connections, generating 2D warps at each stage.

Group Residual Finite Scalar Quantization (GRFSQ)
Our GRFSQ quantizer combines Group Vector Quantization (Group VQ), Residual Vector Quantization (Residual VQ), and Finite Scalar Quantization (FSQ). The GRFSQ process is visually illustrated in Figure 2 (a), with the detailed algorithm presented in Algorithm 1.

Specifically, the input \mathbf{x} is first split into G groups. Each group undergoes R iterations of residual quantization, with FSQ applied to the residual at each step. Similar to the single-codebook VQ (Vector Quantization) method, FSQ projects the continuous representation onto a fixed code-word. However, unlike traditional VQ, FSQ does not involve a codebook lookup process. Instead, each dimension is quantized to a set of fixed levels $[l_1, l_2, \dots, l_d]$, and this process can be optimized using the Straight-Through Estimator (STE). This approach creates an implicit codebook from the product of these level sets, enabling efficient quantization without an explicit codebook.

Finally, the whole process generates $G \times R$ sets of codebook indices \mathbf{I} . The quantized outputs from all groups are merged to form the final quantized vector $\hat{\mathbf{x}}$. These indices \mathbf{I}

Algorithm 1: Group-Residual FSQ

Input: \mathbf{x} , FSQ levels $[l_1, l_2, \dots, l_d]$, number of groups G , number of residual quantizers R
Output: quantized $\hat{\mathbf{x}}$, codebook indices \mathbf{I}
 /* Split into G groups */
 split \mathbf{x} into G groups: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_G$
 $\hat{\mathbf{x}}_g \leftarrow 0$ for $g = 1$ to G
 $\mathbf{I} \leftarrow$ empty tensor of shape $(G, R, *)$
for $g = 1$ to G **do**
 residual $\leftarrow \mathbf{x}_g$
 /* Residual quantizers */
 for $r = 1$ to R **do**
 $z_{\text{quantized}}, \text{indices} \leftarrow$ FSQ (residual)
 $\hat{\mathbf{x}}_g \text{ += } z_{\text{quantized}}$
 residual $\text{ -= } z_{\text{quantized}}$
 $\mathbf{I}[g, r] \leftarrow$ indices
 $\hat{\mathbf{x}} = \text{concat}(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_G)$
return $\hat{\mathbf{x}}, \mathbf{I}$

will serve as target values in the subsequent Coarse-to-Fine Motion Generation process.

This combined strategy of the codebook offers several advantages: it enhances codebook utilization, mitigates codebook collapse without requiring complex auxiliary losses, and provides a flexible framework for capturing multi-scale facial motion features.

Coarse-to-Fine Motion Generation

Our coarse-to-fine motion generation process, illustrated in Figure 2 (b), transforms audio input into facial animation output. This approach employs a BERT (Devlin et al. 2019) model to convert discrete audio features into facial motion tokens, progressively refining from coarse to fine motions.

The process begins with inputs from multiple sources: (1) discrete speech tokens (\mathbf{A}) from a pretrained Speech Tokenizer, (2) additional control information including frame-level head pose (\mathbf{h}), gaze direction (\mathbf{g}), and eye blink (\mathbf{b}), (3) the global feature f_g , and (4) a layer indicator (k). These inputs are combined to form the composite input vector for each residual layer r :

$$\mathbf{C}_r = \{f_g, k, [\mathbf{A}, \mathbf{h}, \mathbf{g}, \mathbf{b}, \mathbf{R}_{r-1}]\}$$

where \mathbf{R}_{r-1} represents the output from the previous iteration (replaced with a zero matrix when $r = 0$), and the brackets represent frame-by-frame concatenation.

The BERT model processes this input to generate facial motion tokens across K residual layers, represented by the ‘Repeat K times’ loop in Figure 2 (b). Each layer refines the motion representation, progressing from coarse to fine details. The Image Renderer then synthesizes the final talking head video from these representations.

Figure 3 illustrates our hybrid facial motion generation matrix, which combines non-autoregressive temporal generation with autoregressive granular refinement. Rows (r_1 to r_4) represent residual codebooks of increasing detail, while t_1 to t_4 indicate the sequential processing steps for each

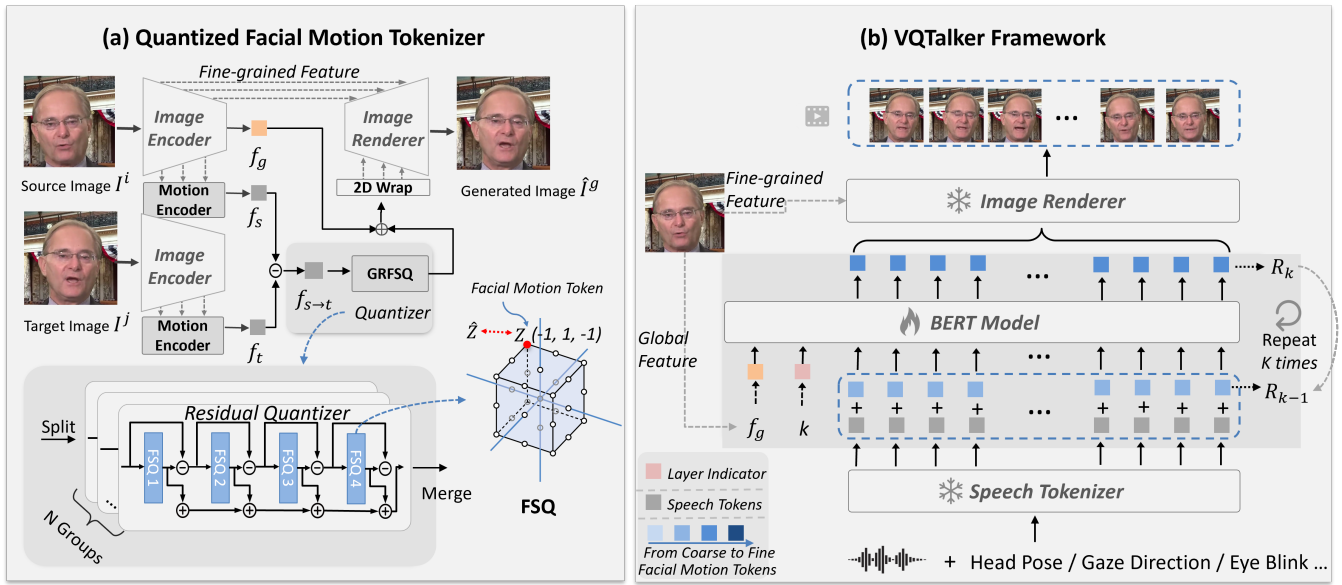


Figure 2: The VQTalker framework consists of two main components: (1) A quantized facial motion codec that learns a universal motion representation through self-supervised learning and Group Residual FSQ (GRFSQ). (2) A coarse-to-fine motion generation process using a BERT model, which takes speech tokens and iteratively generates facial motion tokens across multiple residual layers. The Image Renderer then synthesizes the final talking head video from the generated discrete codewords.

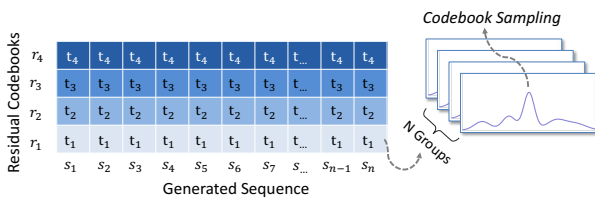


Figure 3: Codebook interleaving patterns. Hybrid facial motion generation matrix. Rows (r_1 to r_4) represent residual codebooks of increasing detail. t_1 to t_4 indicate the sequential processing steps for each residual layer, with each t_i generating the corresponding r_i in one pass.

residual layer. Temporally (horizontally), each sequence s_1, s_2, \dots, s_n is generated simultaneously. Granularly (vertically), generation progresses through four residual layers, with each t_i generating the corresponding r_i in one pass.

At each sequence position and residual layer, we apply group quantization, dividing the representation into N groups. During training, each group is treated as a separate classification task. At inference, we perform argmax sampling for each group. The learning objective is to reconstruct the categories of the $(r + 1)$ -th codebook, conditioned on the input C_r from the r -th layer, by minimizing the negative log-likelihood:

$$L(\mathcal{G}|\mathbf{C}_r) = -\sum_{g=1}^G \sum_{t=1}^T \log p_{\mathcal{G}}(\mathbf{I}_{g,t,r+1}|\mathbf{C}_r)$$

where $\mathbf{I}_{g,t,r+1}$ represents the target codeword index in the

$(r + 1)$ -th layer codebook for the g -th group at time step t . These indices are derived from the GRFSQ process described earlier, serving as the target values for our coarse-to-fine generation. The generation across different residual layers shares the same set of parameters, distinguished by the layer indicator k .

This coarse-to-fine approach allows our model to capture both global facial movements and fine-grained details, resulting in more natural and expressive talking head across diverse linguistic contexts. By using the quantized indices \mathbf{I} as targets, we maintain a direct link between the GRFSQ quantization and the motion generation process, ensuring that the generated motions accurately reflect the discretized facial dynamics captured in the tokenization stage.

Empirically, we observe that different residual layers model distinct aspects of facial motion. The first layer typically captures large-scale movements such as head pose and overall facial orientation. Subsequent layers progressively refine the motion, with middle layers focusing on more localized movements like eyebrow raises, jaw motions, and lip movements. The final few layers capture the most subtle details, including gradually reducing visual jitter. This hierarchical representation allows our model to efficiently capture the full spectrum of facial dynamics, from global head movements to micro-expressions, while also alleviating the complexity of the generation process. As a result, it contributes to the naturalness and expressiveness of the generated animations across various linguistic contexts. For a visual demonstration of this coarse-to-fine progression, we encourage readers to view the accompanying video in the supplementary materials.

Experiment

Experiment Setup

Training Dataset We utilized three publicly available datasets: VoxCeleb (Nagrani, Chung, and Zisserman 2017), HDTF (Zhang et al. 2021), and VFHQ (Xie et al. 2022). To ensure consistency, we re-downloaded the original videos and processed them using a unified approach. This approach included filtering out faces smaller than 512 pixels. Our effort resulted in a dataset comprising around 16k video clips with a total duration of 210 hours. After statistical analysis, **Indo-European** languages (such as English, German, and French) account for around 97% of the final dataset.

Evaluation Dataset To evaluate the effectiveness of non-talking head algorithms on non-Indo-European languages, we compiled the **Multilingual Non-Indo-European Talking Head Evaluation Corpus (MNTE)**. This dataset encompasses Arabic, Swahili, Mandarin, Korean, Japanese, and Turkish, aiming to cover a wide range of commonly used non-Indo-European language families. For each language, we collected 5 talking head video clips, each no shorter than 3 seconds, resulting in a total of 30 videos. To evaluate performance in Indo-European languages and video reconstruction tasks, we use HDTF (Zhang et al. 2021) as our test set, which follows Dinet (Zhang et al. 2023b).

Evaluation Metric We employ a suite of measures to assess the quality and similarity of generated images and videos across different scenarios. These metrics include: Image Similarity Metrics - Structural Similarity Index (SSIM)(Wang et al. 2004) and Learned Perceptual Image Patch Similarity (LPIPS)(Zhang et al. 2018), which quantify structural and perceptual similarity between generated and ground truth images; Distribution-based Metrics - Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD); Facial Fidelity Metrics - Cosine Similarity (CSIM), which evaluates facial similarity, and Landmark Distance (LMD), which measures key facial positioning; Video Quality Metrics - Motion Stability Index (MSI)(Ling et al. 2023), which assesses motion stability; and Image Quality Metrics - Cumulative Probability of Blur Detection (CPBD)(Bohr et al. 2013). To ensure a fair comparison, the resolution is standardized to 256×256 .

Model Configuration For training the face motion tokenizer, our training paradigm is primarily based on LIA (Wang et al. 2022) and extends it to support a resolution of 512×512 pixels. Then, based on it, we add a vector quantization module. In the second stage, we employed a 12-layer BERT (Devlin et al. 2019) network to iteratively generate a four-layer residual codebook for the face tokenizer. The maximum length is 4096. A pre-trained speech tokenizer from CosyVoice (Du et al. 2024) serves as the audio feature encoder, incorporating a downsampling layer to adjust the audio sampling rate from 50 Hz to 25 Hz to synchronize with the video frame rate.

Facial Motion Tokenizer

We conducted a comprehensive evaluation of our approach against several state-of-the-art face reenactment methods (Siarohin et al. 2019; Pang et al. 2023; Tao et al.

Method	Reso.	Bitrate	SSIM \uparrow	LPIPS \downarrow	CSIM \uparrow	CPBD \uparrow
H.264 Codec	512	≈ 347	-	-	-	-
FOMM	256	48	0.775	0.178	0.830	0.339
DPE	256	16	0.861	0.151	0.912	0.354
MTIA	256	48	0.870	0.122	0.929	0.316
Vid2Vid	256	36	0.870	0.115	0.924	0.386
LIA	256	<u>16</u>	0.831	0.137	0.916	0.297
FADM	256	36	0.849	0.147	0.916	0.332
AniTalker	256	<u>16</u>	0.905	0.079	<u>0.927</u>	0.367
LivePortrait	512	≈ 50	0.800	0.134	0.883	0.374
EMOPortrait	512	≈ 102	0.736	0.206	0.657	0.434
VQTalker (Ours)	512	≈ 11	<u>0.874</u>	<u>0.083</u>	0.919	<u>0.421</u>

Table 1: Video Reconstruction Result on HDTF. *Reso.* indicates image resolution (width and height), *Bitrate* is measured in kbps. **Bold** and underlined values represent the best and second-best results respectively.

2022; Wang, Mallya, and Liu 2021; Wang et al. 2022; Zeng et al. 2023; Liu et al. 2024; Chen et al. 2024; Guo et al. 2024) in a video reconstruction scenario. This scenario tests the algorithms’ ability to accurately reconstruct the original video by using a video of the same identity to drive a portrait, with the first frame serving as the source image. All compared methods utilize variations of self-supervised learning. Table 1 presents the results of this evaluation.

Our method, VQTalker, stands out by achieving competitive results while supporting a higher resolution (512×512) and maintaining a significantly lower bitrate. Specifically, our approach employs 12 group layers, 4 residual layers, and 625 codebook entries per group, with a sampling rate of 25 fps, resulting in a total bitrate of approximately 11 kbps (calculated as $4 \times 12 \times \log_2(625) \times 25$). For other methods using continuous variables, we converted them to bits by multiplying by 32-bit float, with all bitrates calculated at 25 fps. For example, LIA, which uses 20 dimensions for intermediate latent, has a bitrate of $20 \times 32 \times 25 = 16000$ (16kbps).

Despite the low bit rate, our method demonstrates strong performance in key metrics. VQTalker achieves the second-highest scores in image structural metrics such as SSIM and LPIPS. Our CSIM is very close to the state-of-the-art, demonstrating robust capabilities in preserving facial identity during reconstruction. In particular, our method achieves the highest CPBD, indicating a better image sharpness compared to other approaches. Although we did not achieve the best results in all metrics, we attribute this to information loss from quantization. For example, subtle position shifts that are imperceptible to the human eye may be captured by these metrics. Compared to traditional video codecs like H.264 (Wiegand et al. 2003), which requires an average bitrate of approximately 347 kbps at 512 resolution, our approach achieves comparable performance while reducing the bitrate to approximately 11 kbps. This suggests that our discrete representation captures facial dynamics more efficiently than the continuous representation. For 512-resolution algorithms, LivePortrait (Guo et al. 2024) requires a bitrate of approximately 50 kbps, but achieves lower performance in most metrics compared to our method. These comparisons highlight our ability to maintain high-quality outputs at higher resolutions while reducing the bandwidth.

Method	HDTF (Indo-European)						MNTE (Non-Indo-European)					
	SSIM↑	CSIM↑	LMD↓	CPBD↑	FID↓	FVD↓	SSIM↑	CSIM↑	LMD↓	CPBD↑	FID↓	FVD↓
SadTalker (Zhang et al. 2023a)	0.510	0.726	0.320	0.391	37.699	526.194	0.456	0.423	0.686	0.216	81.128	564.047
EAT (Gan et al. 2023)	0.433	0.678	0.388	0.334	177.317	667.499	0.394	0.420	0.663	0.224	158.465	540.128
PD-FGC (Wang et al. 2023)	0.325	0.298	0.671	0.264	214.200	846.211	0.357	0.302	0.726	0.212	154.394	624.143
AniTalker (Liu et al. 2024)	0.663	0.709	0.426	0.394	34.705	444.162	0.520	0.490	0.671	0.246	65.780	430.209
EDTalker (Tan et al. 2024a)	<u>0.823</u>	<u>0.881</u>	<u>0.084</u>	0.301	36.437	410.709	0.745	<u>0.802</u>	0.107	0.219	<u>55.070</u>	<u>321.825</u>
EchoMimic (Chen et al. 2024)	0.690	<u>0.881</u>	0.094	0.385	36.354	<u>272.425</u>	0.600	0.737	0.213	0.238	62.185	391.568
VQTalker (Ours)	0.835	0.902	0.067	0.408	28.783	205.471	<u>0.730</u>	0.809	<u>0.146</u>	0.263	43.411	233.586

Table 2: Quantitative comparisons with speech-driven baselines on the HDTF (Indo-European) and MNTE (Non-Indo-European) datasets. **Bold** and underlined values represent the best and second-best results respectively.



Figure 4: Qualitative comparison of speech-driven talking head generation methods across non-Indo-European languages from Arabic to Turkish. Each row shows a source portrait (leftmost) animated by different algorithms to pronounce specific syllables (indicated in red). The upper and lower images in each example demonstrate the transition in lip movements. This visualization highlights our method’s ability to model comprehensive pronunciation features and support multiple languages. To better observe the motion changes and co-articulation capabilities, it is recommended to watch the supplementary video.

Coarse-to-fine Experiment

We evaluate our method against several state-of-the-art speech-driven approaches: SadTalker (Zhang et al. 2023a), EAT (Gan et al. 2023), PD-FGC (Wang et al. 2023), AniTalker (Liu et al. 2024), EDTalker (Tan et al. 2024a), and EchoMimic (Chen et al. 2024). EchoMimic operates at 512×512 resolution, while the other methods work at 256×256 . Our experimental setup involves using the audio from a video to animate its first frame, aiming to recreate the original video as accurately as possible. This evaluation is conducted across both Indo-European and non-Indo-European languages to assess cross-lingual performance.

The quantitative results, presented in Table 2, demonstrate the superior performance of VQTalker. Our method achieves top scores in identity preservation (CSIM), image sharpness (CPBD) and distribution distance (FID and FVD). Additionally, VQTalker maintains competitive results in structural similarity (SSIM) and landmark accuracy (LMD). Notably, our approach exhibits a smaller performance gap between Indo-European and non-Indo-European languages compared to other methods, underscoring its robust cross-lingual generalization. The qualitative results, illustrated in Figure 4, provide visual evidence of VQTalker’s effectiveness in capturing pronunciation actions across diverse non-Indo-European languages. Our method demonstrates notable accuracy in lip movements while consistently preserving identity. This is particularly evident in the natural mouth shapes and fluid transitions observed for various phonemes, such as ‘/ka.e.t.te/’ in Japanese. These examples highlight VQTalker’s proficiency in modeling a wide range of phonetic articulations across different linguistic systems. The visual comparison further emphasizes our model’s ability to adapt to the unique characteristics of each language.

Ablation Study

Method	# G	# R	FSQ	# Codes	Bitrate	CSIM ↑	MSI ↑	Util.(%) ↑
VQ	1	-	-	8196	<1	0.588	0.588	0.35
GVQ	32	-	-	1024	8	0.548	0.052	88.89
RVQ	1	32	-	1024	8	0.561	0.473	50.00
GRVQ	12	4	-	1024	12	0.572	0.359	26.78
GRFSQ	8	4	✓	$5 \times 5 \times 5 \times 5$	≈ 7	0.576	1.016	73.99
GRFSQ	12	2	✓	$5 \times 5 \times 5 \times 5$	≈ 6	0.576	0.616	56.89
GRFSQ	12	4	✓	$4 \times 4 \times 4 \times 4$	9.6	0.570	0.656	59.68
GRFSQ	12	4	✓	$6 \times 6 \times 6 \times 6$	≈ 12	0.571	0.932	65.64
GRFSQ	12	4	✓	$5 \times 5 \times 5 \times 5$	≈ 11	0.571	0.992	78.74

Table 3: Vector Quantization Ablation on HDTF Cross-Identity Driven: # G (groups), # R (residuals), FSQ (Finite Scalar Quantization), # Codes (codebook size / FSQ levels), Util.: Codebook Utilization Rate (code used at least once)

Codebook Design Table 3 compares different codebook designs in cross-identity-driven scenarios. We focus on cross-identity scenarios here as they reveal more significant differences between methods, whereas self-driven scenarios show less pronounced variations. Classic designs (VQ, GVQ, RVQ, GRVQ) generally show lower MSI scores, indicating increased jitter on the generated result, suggesting limited model capacity in modeling complex facial dy-

namics. Our GRFSQ variations demonstrate that balancing group number, residual layers, and FSQ levels is crucial. The optimal configuration (12 groups, 4 residuals, $5 \times 5 \times 5 \times 5$ FSQ levels) achieves a good trade-off between CSIM, MSI, and codebook utilization. Notably, GRFSQ outperforms GRVQ in MSI (0.992 vs 0.359) and utilization rate (78.74% vs 26.78%), suggesting FSQ’s structure allows for more efficient use of the representation space, potentially leading to improved motion modeling and stability.

Method	Audio Feature	Motion Feature	SSIM ↑	CSIM ↑	LPIPS ↓	CPBD ↑
C-C	Whisper	Vector	0.795	0.889	0.101	0.410
D-C	CosyVoice	Vector	0.769	0.882	0.115	0.410
C-D	Whisper	GRFSQ	<u>0.821</u>	<u>0.890</u>	<u>0.085</u>	<u>0.408</u>
D-D	VQ-Wav2vec	GRFSQ	0.804	0.887	0.089	<u>0.408</u>
D-D	CosyVoice	GRFSQ	0.835	0.902	0.067	<u>0.408</u>

Table 4: Ablation on Discrete (D) and Continuous (C) Representation. X-Y format: X for input (Audio Feature), Y for output (Motion Feature).

Discrete vs. Continuous Representation Table 4 compares discrete and continuous representations for audio and motion features. For continuous audio representation, we employed the multilingual Whisper Large v2 (Radford et al. 2023) model, ensuring a fair comparison across different languages. Discrete representations consistently outperform continuous ones, especially when used for both input and output (D-D). Our proposed method (CosyVoice for audio, GRFSQ for motion) achieves the best overall performance, with notable improvements in SSIM (0.835) and LPIPS (0.067). This suggests that discrete representations can more effectively capture relevant information for both audio and motion features, even when compared to advanced multilingual models like Whisper. Among speech tokenizers, CosyVoice outperforms VQ-Wav2vec (Baevski, Schneider, and Auli 2020), which may be attributed to its more compact and efficient codebook design. These results underscore the potential of leveraging discrete representations in both audio and motion domains for multilingual talking head generation tasks.

Conclusion

We introduced VQTalker, a framework designed for multilingual talking head generation, utilizing discretized facial motion representations. The Group Residual Finite Scalar Quantization (GRFSQ) method within VQTalker successfully trades off low bitrate with high-quality facial animation. VQTalker particularly excels in speech-driven scenarios, offering robust performance across non-Indo-European languages. The coarse-to-fine generation approach contributes to improved temporal consistency and natural lip synchronization, even in diverse linguistic contexts. These results underscore the effectiveness of discretized representations in capturing complex facial dynamics while maintaining efficiency, advancing the development of more inclusive and versatile talking head technologies.

Acknowledgments

This work has been supported by the China NSFC Project (No. 92370206), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and the Key Research and Development Program of Jiangsu Province, China (Grant No. BE2022059).

References

- Baevski, A.; Schneider, S.; and Auli, M. 2020. vq-wav2vec: Self-supervised learning of discrete speech representations. *International Conference on Learning Representations (ICLR)*.
- Bohr, P.; Gargote, R.; Vhorkate, R.; Yawle, R.; and Bairagi, V. 2013. A no reference image blur detection using cumulative probability blur detection (cpbd) metric. *International Journal of Science and Modern Engineering*, 1(5).
- Chen, L.-W.; Watanabe, S.; and Rudnicky, A. 2023. A vector quantized approach for text to speech synthesis on real-world spontaneous speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 12644–12652.
- Chen, Z.; Cao, J.; Chen, Z.; Li, Y.; and Ma, C. 2024. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions. *arXiv preprint arXiv:2407.08136*.
- Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.
- Drobyshev, N.; Casademunt, A. B.; Vougioukas, K.; Landgraf, Z.; Petridis, S.; and Pantic, M. 2024. EMOPortraits: Emotion-enhanced Multimodal One-shot Head Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8498–8507.
- Drobyshev, N.; Chelishev, J.; Khakhulin, T.; Ivakhnenko, A.; Lempitsky, V.; and Zakharov, E. 2022. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2663–2671.
- Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; et al. 2024. CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. *arXiv preprint arXiv:2407.05407*.
- Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.
- Fabian, M.; David, M.; Eirikur, A.; and Michael, T. 2024. Finite Scalar Quantization: VQ-VAE Made Simple. *International Conference on Learning Representations (ICLR)*.
- Gan, Y.; Yang, Z.; Yue, X.; Sun, L.; and Yang, Y. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22634–22645.
- Guo, J.; Zhang, D.; Liu, X.; Zhong, Z.; Zhang, Y.; Wan, P.; and Zhang, D. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168*.
- He, T.; Guo, J.; Yu, R.; Wang, Y.; Zhu, J.; An, K.; Li, L.; Tan, X.; Wang, C.; Hu, H.; Wu, H.; Zhao, S.; and Bian, J. 2024. GAIA: Zero-shot Talking Avatar Generation. *The International Conference on Learning Representations (ICLR)*.
- Huang, H.; Wu, Z.; Kang, S.; Dai, D.; Jia, J.; Fu, T.; Tuo, D.; Lei, G.; Liu, P.; Su, D.; et al. 2021. Speaker independent and multilingual/mixlingual speech-driven talking head generation using phonetic posteriorgrams. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1433–1437. IEEE.
- Ling, J.; Tan, X.; Chen, L.; Li, R.; Zhang, Y.; Zhao, S.; and Song, L. 2023. Stableface: Analyzing and improving motion stability for talking face generation. *IEEE Journal of Selected Topics in Signal Processing*.
- Liu, T.; Chen, F.; Fan, S.; Du, C.; Chen, Q.; Chen, X.; and Yu, K. 2024. AniTalker: Animate Vivid and Diverse Talking Faces through Identity-Decoupled Facial Motion Encoding. In *Proceedings of the 32th ACM International Conference on Multimedia (ACM MM)*.
- Ma, Y.; Wang, S.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; Deng, Z.; and Yu, X. 2023. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1896–1904.
- McGurk, H.; and MacDonald, J. 1976. Hearing lips and seeing voices. *Nature*.
- Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.
- Pang, Y.; Zhang, Y.; Quan, W.; Fan, Y.; Cun, X.; Shan, Y.; and Yan, D.-m. 2023. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 427–436.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.
- Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in neural information processing systems*, 32.
- Song, H.-K.; Woo, S. H.; Lee, J.; Yang, S.; Cho, H.; Lee, Y.; Choi, D.; and Kim, K.-w. 2022. Talking face generation with multilingual tts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21425–21430.

- Sung-Bin, K.; Chae-Yeon, L.; Son, G.; Hyun-Bin, O.; Ju, J.; Nam, S.; and Oh, T.-H. 2024. MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset. *arXiv preprint arXiv:2406.14272*.
- Tan, S.; Ji, B.; Bi, M.; and Pan, Y. 2024a. EDTalk: Efficient Disentanglement for Emotional Talking Head Synthesis. *Proceedings of the European conference on computer vision (ECCV)*.
- Tan, S.; Ji, B.; Ding, Y.; and Pan, Y. 2024b. Say anything with any style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5088–5096.
- Tan, S.; Ji, B.; and Pan, Y. 2024. FlowVQTalker: High-Quality Emotional Talking Face Generation through Normalizing Flow and Quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26317–26327.
- Tao, J.; Wang, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Motion Transformer for Unsupervised Image Animation. In *European Conference on Computer Vision (ECCV)*, 702–719. Springer.
- Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*.
- Wang, D.; Deng, Y.; Yin, Z.; Shum, H.-Y.; and Wang, B. 2023. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17979–17989.
- Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision (ECCV)*, 700–717. Springer.
- Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 10039–10049.
- Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent image animator: Learning to animate images via latent space navigation. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.
- Wiegand, T.; Sullivan, G. J.; Bjontegaard, G.; and Luthra, A. 2003. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7): 560–576.
- Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 657–666.
- Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023a. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12780–12790.
- Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023b. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12780–12790.
- Xu, S.; Chen, G.; Guo, Y.-X.; Yang, J.; Li, C.; Zang, Z.; Zhang, Y.; Tong, X.; and Guo, B. 2024. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*.
- Yang, Y.; Shillingford, B.; Assael, Y.; Wang, M.; Liu, W.; Chen, Y.; Zhang, Y.; Sezener, E.; Cobo, L. C.; Denil, M.; et al. 2020. Large-scale multilingual audio visual dubbing. *arXiv preprint arXiv:2011.03530*.
- Yu, Q.; Weber, M.; Deng, X.; Shen, X.; Cremers, D.; and Chen, L.-C. 2024. An Image is Worth 32 Tokens for Reconstruction and Generation. *arXiv preprint arXiv:2406.07550*.
- Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 30: 495–507.
- Zeng, B.; Liu, X.; Gao, S.; Liu, B.; Li, H.; Liu, J.; and Zhang, B. 2023. Face Animation with an Attribute-Guided Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 628–637.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.
- Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023a. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8652–8661.
- Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhang, Z.; et al. 2023b. DINet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video. *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*.
- Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision (ECCV)*, 650–667. Springer.