HUMAN-ROBOT INTERACTION

Learning realistic lip motions for humanoid face robots

Yuhang Hu[1]*†, Jiong Lin[1], Judah Allen Goldfeder[2], Philippe M. Wyder[1]‡, Yifeng Cao[3], Steven Tian[1], Yunzhe Wang[2]§, Jingran Wang[1], Mengmeng Wang[1], Jie Zeng[1], Cameron Mehlman[1]¶, Yingke Wang[2]#, Delin Zeng[1], Boyuan Chen[4], Hod Lipson[1,5]*

Lip motion represents outsized importance in human communication, capturing nearly half of our visual attention during conversation. Yet anthropomorphic robots often fail to achieve lip-audio synchronization, resulting in clumsy and lifeless lip behaviors. Two fundamental barriers underlay this challenge. First, robotic lips typically lack the mechanical complexity required to reproduce nuanced human mouth movements; second, existing synchronization methods depend on manually predefined movements and rules, restricting adaptability and realism. Here, we present a humanoid robot face designed to overcome these limitations, featuring soft silicone lips actuated by a 10–degree-of-freedom mechanism. To achieve lip synchronization without predefined movements, we used a self-supervised learning pipeline based on a variational autoencoder (VAE) combined with a facial action transformer, enabling the robot to autonomously infer more realistic lip trajectories directly from speech audio. Our experimental results suggest that this method outperforms simple heuristics like amplitude-based baselines in achieving more visually coherent lip-audio synchronization. Furthermore, the learned synchronization successfully generalizes across multiple linguistic contexts, enabling robot speech articulation in 10 languages unseen during training.

INTRODUCTION

Imagine sitting across from a robot that can hold a conversation, its lips moving in perfect harmony with its words. You would not just hear its voice, you would see it speak, just like a human. This blend of auditory and visual cues is how we naturally engage with each other, and it is why incongruence in lip and audio synchronization feels uncanny and unsettling to us (1–4).

Proper lip movements are also crucial for understanding content. A combination of auditory and visual speech recognition can be more accurate than just receiving one or the other (5–7). Humans can easily and keenly perceive when visual cues do not match auditory cues. Studies have shown that in noisy environments, observers increasingly rely on visual cues from the speaker's lips, with fixation on the mouth region rising substantially under such conditions, reaching about 50 to 55% of gaze time (8).

We suggest that for humans to be more willing to communicate with anthropomorphic robots, it is essential that such robots have the ability to synchronize lips and speech with humans. Without this ability, even a robot with an advanced humanoid appearance will appear lifeless, resulting in the notorious uncanny valley effect (9, 10), and people may quickly lose interest or trust in the interaction.

Over the years, researchers have shown that robots with a humanlike appearance are one of the ideal human-robot interaction platforms because they can convey emotions and deeper cues through facial expressions, allowing people to engage more in emotional communication with robots (11–13). Achieving real-time, realistic lip-audio synchronization in humanoid robots is a long-standing challenge and has been addressed from various perspectives in prior work. For example, Ishi et al. (14) proposed a formant-based lip motion generation method in teleoperated humanoid robots, mapping acoustic speech features to predefined articulatory movements. Strathearn and Ma (15) developed a robotic articulation system using a precisely engineered mechanical mouth driven by a phoneme-to-motion control scheme. These efforts demonstrated that accurate lip synchronization is achievable through heuristic or rule-based control methods. However, such approaches often require extensive manual tuning and may lack flexibility for expressive, speaker-dependent, or multilingual speech.

Some previous work used the categorization of phonetic symbols for lip synchronization and designed predefined lip movements for each category, including motion trajectories and duration time (16–18). Such methods, although simple and transparent, have several challenges. First, the design of each lip movement is time-consuming and labor-intensive. Second, the lip movement speed of the keywords needs to be adjusted during the motor execution, which requires understanding the speech content and transferring it into text to effectively achieve more realistic lip synchronization. Similar to attempts in manual design of robot locomotion, manual tuning of motion primitives has limits, but it fails to improve with more data and experience.

Computer graphics and deep learning researchers have already recognized that lip-sync technology can be better achieved by learning directly from large-scale speech visual-audio datasets with end-to-end neural network models. Prajwal's Wav2Lip model (19) used a generative adversarial network architecture with loss from a pretrained lip-sync expert. Lahiri et al. (20) used a video-based learning framework to animate three-dimensional (3D) talking faces from audio. Digital-only pipelines, such as Voice Operated Character Animation (VOCA), Meshtalk, and Codetalker (21–23),

[1]Department of Mechanical Engineering, Columbia University, New York, NY, USA. [2]Department of Computer Science, Columbia University, New York, NY, USA. [3]Department of Electrical Engineering, Columbia University, New York, NY, USA. [4]Department of Mechanical Engineering & Materials Science, Duke University, Durham, NC, USA. [5]Data Science Institute, Columbia University, New York, NY, USA.
*Corresponding author. Email: yuhang.hu@columbia.edu (Y.H.); hod.lipson@columbia.edu (H.L.)
†Present address: AheadForm, Inc., New York, NY, USA.
‡Present address: Distyl AI, New York, NY, USA.
§Present address: USC Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA.
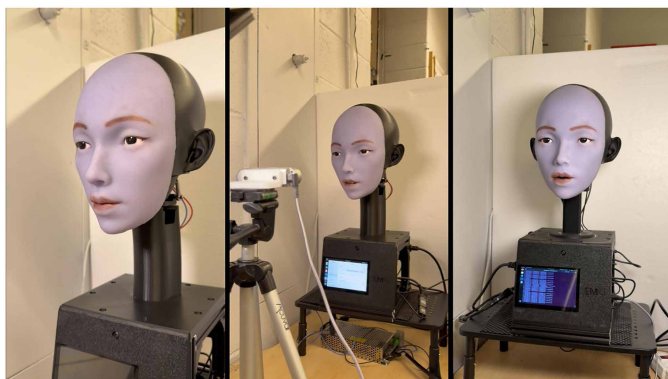¶Present address: Department of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY, USA.
#Present address: Department of Computer Science, Stanford University, Stanford, CA, USA.

learned to map speech to blend shape coefficients of a 3D morphable face. End-to-end models that combine speech and identity features use recurrent structures. These models capture frame-to-frame dynamics and contextual information (*24*). Some recent implementations separate identity from speech information, allowing them to generate speech for different individuals with high fidelity (*25*, *26*). These models can capture context and adapt to different speech conditions, producing lifelike visual representations of speaking. Although highly effective for avatars, these methods assume a differentiable mesh with no actuation limits. In contrast, our robot must contend with servo inertia, elastic skin mechanics, and collision bounds.

Unlike digital avatars, real robotic faces are constrained by mechanical constraints. Motors, servos, and physical linkages dictate the range of motion and the speed at which lip movements can be executed. This complexity is compounded by the nonlinear kinetics and dynamics of the elastic skin and lips tugging back at motors and cladding the face, making some motions kinematically difficult or impossible. This complexity introduces new challenges: How do you ensure that the generated motor commands translate smoothly into physical actions? How do you account for the potential lag between command and execution or the mechanical limitations that might prevent perfect replication of synthesized movements? These are questions that digital models do not face but are critical for real robots to achieve less robotic-looking lip motion. To bridge this complexity, we used a learned self-model (*27*–*30*).

Representation learning aims to find compact, informative representations of input data that can generalize across different domains (*31*–*35*). Our goal is to bridge the distribution gap between 2D video outputs (where lips are merely pixels) and a real, mechanically constrained robot face (where lips are physical actuators and elastomeric skin). By integrating a variational autoencoder (VAE) and a learned robot facial self-model, we aim to transform synthesized video input into smooth and lifelike motor commands that ensure synchronized speech and intuitive-looking interactions (*36*–*39*). Here, we seek to move beyond static, preprogrammed actions and instead leverage a data-driven learning framework that adapts to the complexities of real-world robotic operation, generalizing to diverse speech inputs while capturing audio-driven articulation, thus setting the stage for more fluid and human-like communication with robots (Movie 1).

Movie 1. Overview of learning realistic lip motions for humanoid face robots. The song sung by the robot in movie S5 was generated using the Suno platform.

RESULTS
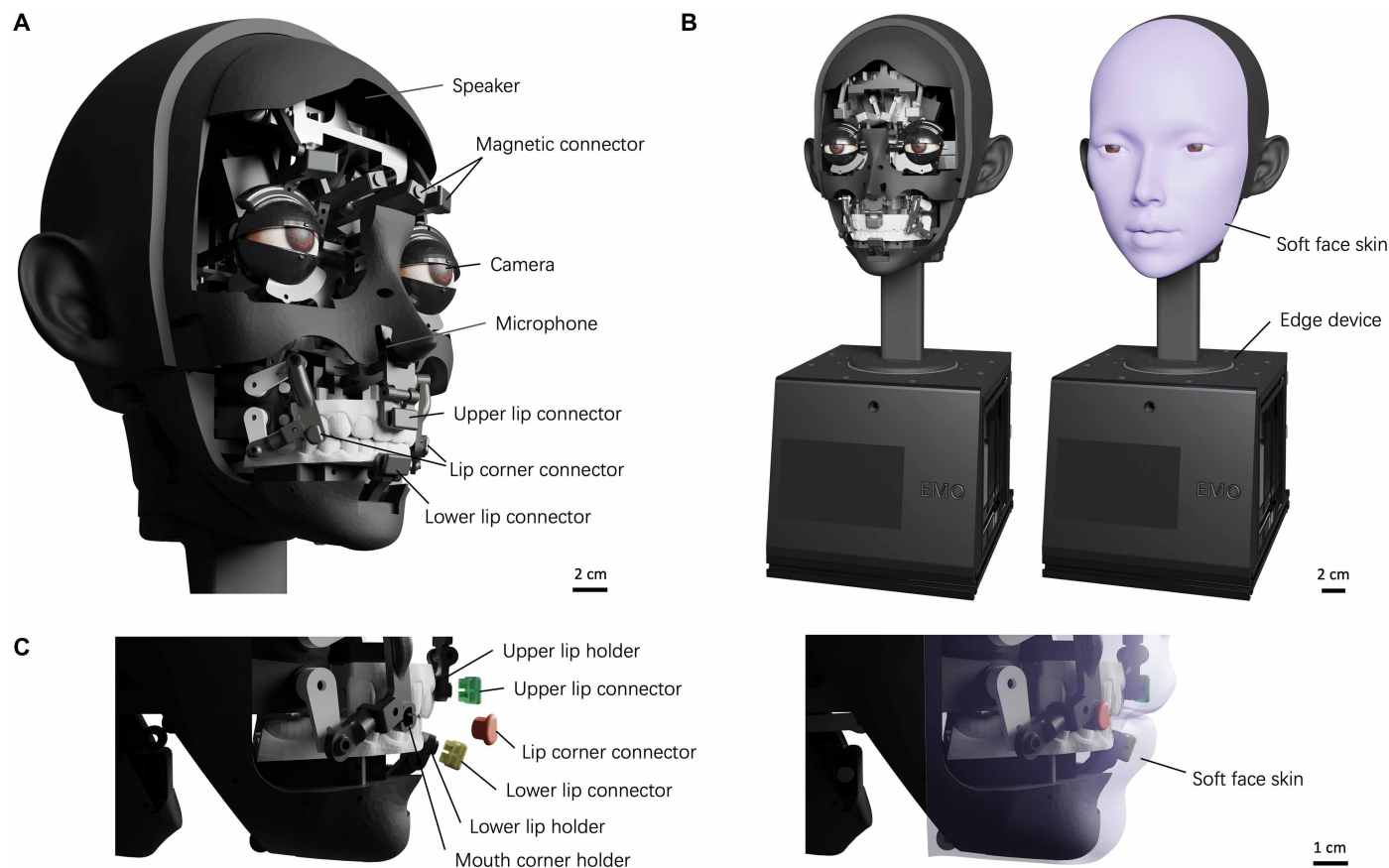### Design of a face with realistic kinematics

Our face robot leverages servo motors, soft silicone face skin, and linkage mechanisms with magnetic connectors to enable lifelike lip movements and real-time interaction (Fig. 1A). This section provides an overview of the components and mechanisms used to overcome the limitations of traditional face robots, which often struggle to synchronize their lips while speaking because of limitations in degrees of freedom and cable-driven designs (*36*, *40*).

The robot's design features a high–degree of freedom (DOF) lip actuation mechanism, offering 10 DOFs: two pairs for the lip corners, three for the upper lip, one for the jaw, and two for the lower lip. The lip corners are controlled by two stacked motors, forming a 2D movement space, allowing both retraction and outward protrusion. This configuration enables complex expressions, such as lip puckering, and provides the ability to form tightly sealed mouth shapes, which are essential for realistic lip motion during speech.
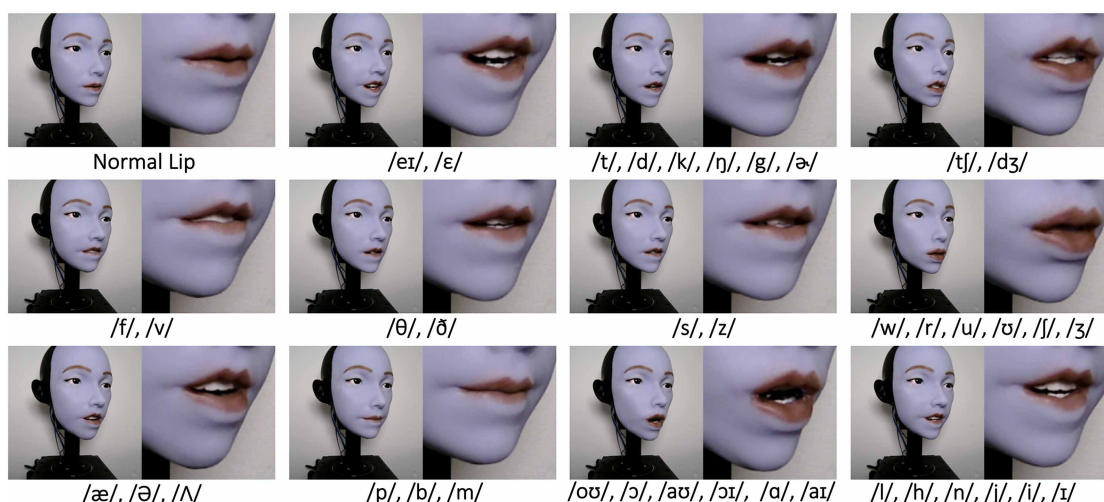
The upper and lower lips are independently actuated in the vertical direction. The upper lip connector turns outward as it descends, imitating the movement of the human upper lip that makes a pout, such as when the mouth makes a "w," "r," or "u" sound, as shown in Fig. 2. Likewise, when the lower lip elevates, its underactuated rotational axis pivots outward, adapting to the upward motion and maintaining a compliant interface with the soft lip. Because these actuators can both push and pull the flexible lip skin, they overcome a fundamental limitation of traditional cable-driven mechanisms, which rely solely on pulling movements.

Our design incorporates magnetic quick-release connectors that align the soft silicone skin precisely with the underlying mechanical infrastructure (Fig. 1B). Each connector is attached to the face skin via super glue and can be easily detached from the four mechanical holders. The modularity of the magnetic connectors facilitates easy skin replacement and maintenance. Unlike other rigid-body robots and serially structured robotic arms, the face robot with soft materials requires constant iteration and position correction during design to form a more realistic mouth shape. Therefore, the quick-release structure facilitates rapid iteration of the design. In contrast, traditional cable-driven facial robots need to calibrate the zero-point position of the pull cord, which results in low iteration efficiency. To enable real-time interaction, the facial robot incorporates high-resolution RGB (red, green, blue) cameras embedded within the eyeballs, providing advanced visual perception and gaze tracking. A microphone and speaker allow the robot to achieve conversational capabilities. The motor control is processed on edge computing devices housed in the robot's base, ensuring low-latency responses and seamless interaction with users.

Our lip mechanism was meticulously designed to cover 24 consonants and 16 vowels, as demonstrated in Fig. 2. Although languages vary in the number of phonemes, English contains approximately 37 to 41 phonemes, depending on the dialect and analysis, with the global average across languages being around 30 (*41*). We classified the robot's viseme into 12 speech-relevant categories, each corresponding to specific phonemes based on typical human lip motions (*42*). These shapes include exposing the upper teeth for sounds like /h/, /l/, and /n/; biting the lower lip for /f/ and /v/; and forming a pouting shape for /w/ and /r/. Unlike traditional face robots that are limited to basic mouth opening and closing movements, these fundamental shapes serve as building blocks for accurate lip synchronization, enabling

Fig. 1. Robot head design featuring advanced mechanical articulation. (**A**) Overview of the facial robot design, highlighting key components for human-robot interaction, including the speaker, microphone, high-resolution camera modules, and magnetic quick-release connectors that secure the soft silicone face skin. The connectors allow for precise alignment and enable both pushing and pulling motions of the skin, facilitating complex lip movements essential for speech articulation. (**B**) The external appearance of the humanoid robot with soft silicone skin. An Edge computing device is housed in the base. (**C**) Detailed view of the lip actuation system, showing the upper, lower, and corner lip connectors, each attached to the corresponding lip holders. The soft, replaceable face skin is secured using magnetic connectors and can be easily detached for maintenance or customization.

Fig. 2. Lip pronunciation movements of the face robot and corresponding phonetic symbols. The robot demonstrates its ability to reproduce key English phonetic symbols, such as plosives (/p/ and /b/), bilabials (/m/), and rounded vowels (/u/ and /o/). Each frame captures the typical lip movements achieved through independent control of the upper, lower, and corner lips. The results are the basis for the robot's ability to produce correct lip alignment when speaking.

fluid human-robot interaction. A detailed demonstration of data collection is available in movie S1.

Demonstrating lip sync through word pronunciation

This section provides an overview of the process used to generate motor commands for the robot's lip synchronization, as shown in Fig. 3. The pipeline incorporated data collection, training, and deployment phases that transformed text inputs into synchronized speech and corresponding lip movements. Detailed information about the model training process can be found in Materials and Methods.

The initial phase involved collecting video footage of the real robot performing various lip movements. These included basic motions such as opening, closing, protruding, and retracting the lips, as well as combinations of movements that mimic the shapes formed during various phonetic articulations (rounded vowels, plosives, and fricatives). As the robot executed each movement, a front-facing camera recorded video footage of its lips. These videos were paired with corresponding speech-relevant motor commands ($A_0, A_1, \ldots, A_t$), which represent the positions and movements of the robot's actuators during speech. These data were used to train a VAE, which encoded synthesized robot video frames into latent vectors that capture the essential features of the real robot's facial movements.

The deployment phase starts with text input. In this section, we used a list of words only, but sentences can be generated by systems such as ChatGPT. This text was converted to audio using a text-to-speech (TTS) system and paired with a synthesized video created by Wav2Lip (43). The synthesized video was processed by the encoder of the trained VAE to produce latent vectors ($L'_0, L'_1, \ldots, L'_t$). These latent vectors served as a reference for generating motor commands.
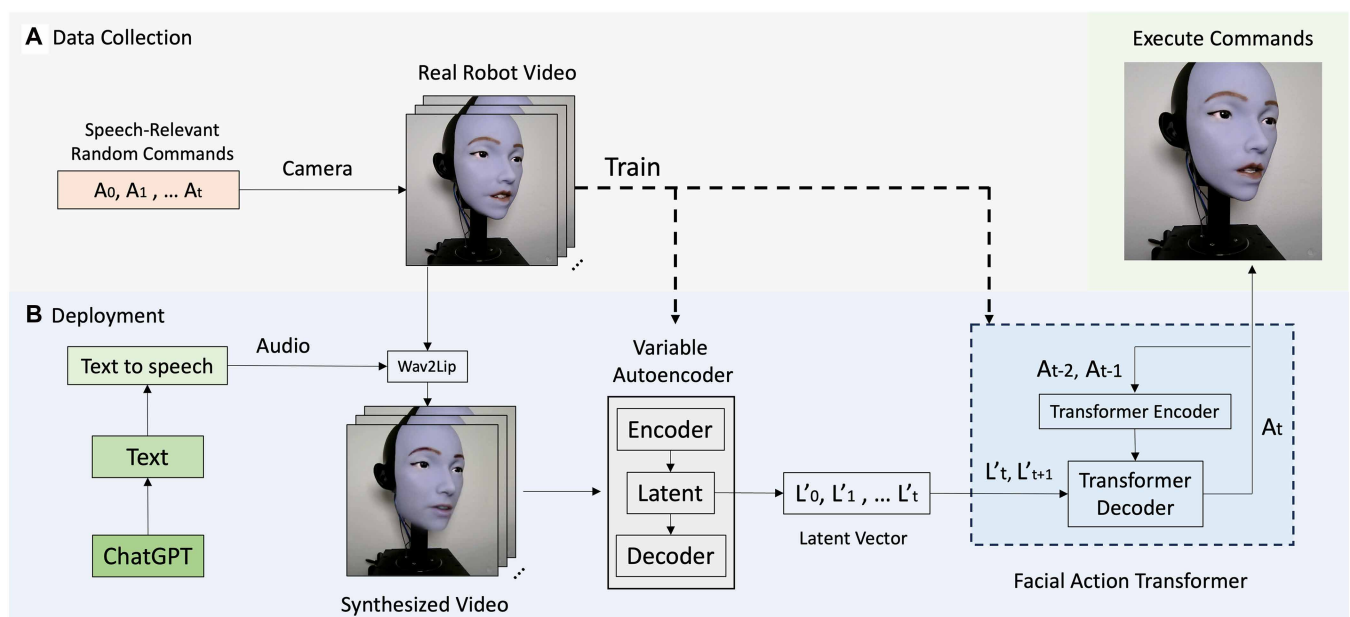
The facial action transformer (FAT) was used to produce smooth and continuous motor commands based on these latent vectors. The transformer encoder takes into account previous motor commands (A_{t-2}, A_{t-1}) to ensure temporal consistency, and the transformer decoder predicts the future motor command (A_t, A_{t+1}) using the previous latent vectors. This prediction allows the robot to replicate the synthesized lip shapes and execute them in real-time.

The result (Fig. 4) showcases the robot's ability to synchronize its lip movements with audio input across various words. For words involving bilabial sounds, such as /m/ in "grandma" or /b/ in "blue," the robot accurately forms the necessary closed-lip shape. This precise closure is vital, because any deviation, such as partially open lips, would be quickly detected by human observers, leading to discomfort or even the uncanny valley effect. In movie S2, we demonstrate the robot's smooth and accurate transitions between different sounds, such as from bilabial to vowel sounds in words like "between" and "bat." Thus, the generated motor commands effectively execute these transitions without visible lags or abrupt changes, which is crucial for maintaining the fluidity of speech. In words that include elongated vowels, such as "father" and "spa," the robot maintains the appropriate lip shape consistently over time. This ability to sustain specific shapes without jitter indicates the effectiveness of the synthesized image pipeline in generating stable commands for continuous motion. The precise formation of lip shapes and their synchronization with audio prevent common issues such as unrealistic movements or delays, which could otherwise trigger discomfort or disengagement. The smooth operation enhances the overall human-robot interaction, making the robot more relatable and engaging.

Evaluation of lip synchronization in continuous speech

This section presents a quantitative evaluation of the proposed lip synchronization method compared with five baseline approaches. The evaluation measures the mean squared error (MSE) between the latent vectors of the synthesized images and the latent vectors of real robot images. This comparison allows us to assess how closely

Fig. 3. Self-supervised learning framework for robotic lip synchronization. (**A**) The data collection phase involves the robot autonomously generating a dataset through speech-relevant random commands, capturing a wide array of lip movements with a side-view camera for 3D lip shape data. (**B**) The deployment process starts with text inputs from ChatGPT that are converted to audio, then synthesizing the robot videos. The real robot video and commands are used to train the Robot Inverse Transformer, which consists of an encoder and decoder, to produce smooth and accurate motor commands for real robot execution.

Fig. 4. Robot lip movements across different phonetic contexts. The figure presents synthesized predictions of the robot's lip shapes for various words alongside real-world robotic performance, illustrating how the learned model generalizes from simulation to physical execution.

the robot's movements align with the ideal (synthesized) visual output. A lower MSE indicates a better match, meaning that the real robot's lip motions closely resemble those in the synthesized videos. The results were validated across three distinct test sentences, and qualitative results are provided in movie S3. The five baselines used for comparison are described as follows. The first baseline, nearest-neighbor (NN) landmarks, used Mediapipe (*44*) to compute distances between the synthesized images and a dataset of real robot images. The closest matching image from the dataset is selected for each frame, representing a feature-based matching approach. The second baseline, amplitude baseline (audio-amplitude jaw motion), replicates the traditional approach used in many face robots, where the jaw opens and closes on the basis of the amplitude of the audio wave. It is simple and lacks the nuances needed for complex speech articulation. The third baseline introduces a 0.033-s temporal shift between the synthesized and real outputs to assess the system's sensitivity to minor temporal misalignments, simulating potential frame drift. The fourth baseline applies a larger temporal shift of 0.5 s to test the effects of substantial desynchronization and to illustrate how severe timing errors degrade alignment quality. The fifth baseline, random selection, assigns random motor commands without reference to the input audio or synthesized visual

features, providing a control condition that reflects the absence of meaningful synchronization.

Our method consistently achieved the lowest MSE across all three test sentences generated by ChatGPT (provided in Materials and Methods), with values of 0.0140, 0.0118, and 0.0136, respectively (Table 1). Our method outperformed the baseline approaches across all three test sentences, demonstrating the effectiveness of leveraging a VAE and a FAT for accurate, real-time lip synchronization. Below, we provide a deeper analysis of the results.

The NN landmarks baseline struggles to match the performance of our method because of the limitations inherent in leveraging facial landmark detection algorithms. Facial landmark detection models extract only the contours or shape of the lips and fail to capture finer details of lip motion. For example, whether the lips are open with teeth exposed or open without showing teeth appears similar to the landmark-based detection given that both scenarios share roughly the same lip shape. However, these differences are crucial for human perception, because they convey subtle variations in phonetic sounds, such as /f/ versus /a/. This lack of detailed motion information reduces the ability of NN landmarks to accurately reflect real speech dynamics, resulting in poorer performance and higher MSE values.

Table 1. MSE comparison of our method and baseline approaches across three test sentences. Lmks, landmarks; BL, baseline; Std., standard deviation; min., minimum; max., maximum.

Metric	Our method	NN Lmks BL	Amplitude BL	Shift 0.033 s	Shift 0.5 s	Random selection
			Test sentence 1 (327 frames)			
Mean	0.0140	0.4014	0.6265	0.0494	0.3606	0.5366
Std.	0.0091	0.2742	0.3018	0.0698	0.2908	0.3899
Min.	0.0025	0.0109	0.1560	0.0032	0.0116	0.0298
Max.	0.0539	1.4756	1.9580	0.6567	1.2822	2.7344
			Sentence 2 (702 frames)			
Mean	0.0118	0.3711	0.7637	0.0563	0.3018	0.5493
Std.	0.0059	0.2544	0.2382	0.0711	0.2371	0.3904
Min.	0.0024	0.0172	0.1587	0.0024	0.0078	0.0297
Max.	0.0411	1.3760	1.7051	0.5513	1.2256	2.4824
			Sentence 3 (444 frames)			
Mean	0.0136	0.3896	0.8276	0.0519	0.2966	0.5771
Std.	0.0062	0.2527	0.2583	0.0647	0.2263	0.3904
Min.	0.0036	0.0158	0.2642	0.0057	0.0068	0.0362
Max.	0.0408	1.4316	1.7393	0.5845	1.1162	2.4531

The amplitude baseline maintains basic alignment between audio amplitude and jaw motion. This method only controls 1 DOF, the up-and-down movement of the jaw, and therefore cannot replicate the complexity of speech-related facial movements, such as lip rounding or corner retraction, which are essential for more realistic lip synchronization. The amplitude baseline performs worse than random selection, with higher MSE values in all three sentences. This finding reinforces the inadequacy of simple audio amplitude-driven lip movements.

Although random selection lacks any deliberate synchronization, it still draws from a real robot dataset containing 20,000 samples distributed according to typical speech patterns. This suggests that speech-relevant movements are embedded within the dataset, even when chosen at random, giving this baseline a slight advantage over the amplitude baseline. The distributional nature of the random selection baseline highlights the importance of training on large, representative datasets. It shows that even without sophisticated algorithms, exposure to diverse, speech-relevant movements improves performance compared with simplistic, rule-based systems like the amplitude baseline.

The 0.33- and 0.5-s shift baselines reveal the sensitivity of lip synchronization to temporal alignment. Even a 0.33-s shift introduces noticeable errors, demonstrating that precise synchronization is crucial for achieving more intuitive interactions. The 0.5-s shift further emphasizes the importance of real-time audio-visual coordination, given that larger temporal misalignments disrupt the continuity of movements and lead to higher MSE.
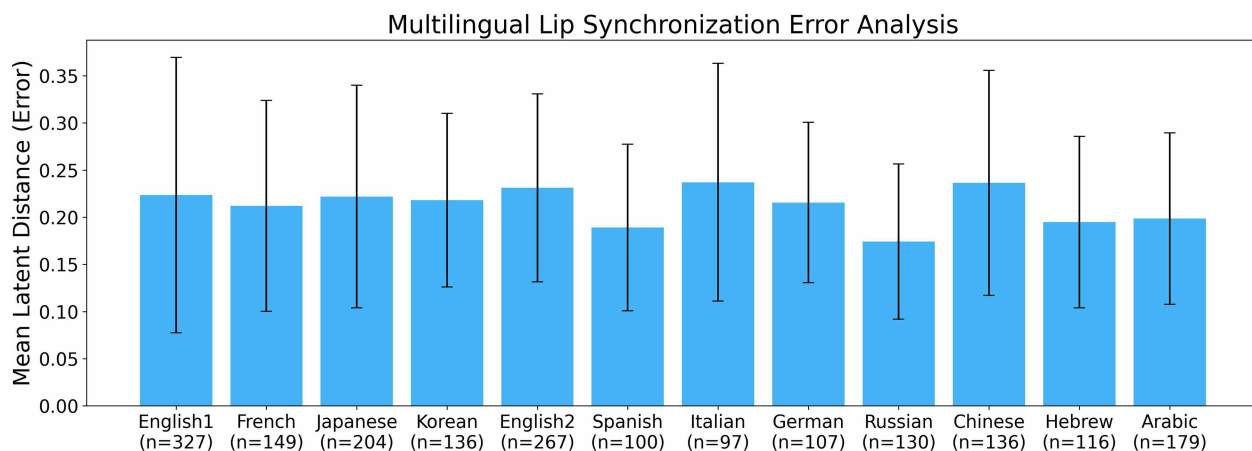
To evaluate whether the proposed lip movements represent an improvement, we conducted a survey comparing our method with the first two baselines. The results show a higher preference for our method (62.5%, $P < 0.0001$). Full details are provided in Supplementary Methods.

Multilingual processing capability
To evaluate the adaptability of our self-supervised lip synchronization framework, we tested the robot with audio inputs from 11 languages: English, French, Japanese, Korean, Spanish, Italian, German, Russian, Chinese, Hebrew, and Arabic. This experiment assessed whether the system could maintain accurate lip synchronization across different linguistic and phonetic contexts. For quantitative evaluation, we used latent distance metrics to measure the synchronization accuracy. The latent distances were computed between the real-robot lip motion videos and the synthesized reference videos. The MSEs and their SDs for all tested languages are plotted in Fig. 5. The mean error values for non-English languages, including languages with different phonetic structures, fell within the range of the English1 (English in female voice) error bars. This consistency confirms that the system can generalize well, maintaining synchronization accuracy even with diverse phonetic challenges. Languages with more complex or nuanced phonetic variations, such as Russian and Chinese, showed slightly higher variability but still remained within an acceptable range. This finding suggests that, although the model is inherently robust, certain languages may challenge it more because of the intricacies of their phonetic and articulatory properties. The robot's consistent performance across languages, as compared to the English1 baseline, implies that even if training data primarily consist of English audio, the model can generalize well to other languages. This highlights an efficient method of data collection, where collecting and training on a single, dominant language lip motion can still yield good performance across multiple linguistic contexts. The result of English2 (an older male voice) demonstrates that the system could process different voice tones within the same language without notable performance drops. This adaptability to different voices within the same language demonstrates the system's potential for deployment in environments where the robot may need to interact with individuals with varied speech patterns, accents, and tones. It demonstrates the system's robustness and flexibility in real-world applications. A demonstration video is provided in movie S4, showcasing the robot speaking in all 11 languages.

The ability to generalize across languages with diverse phonetic and articulatory requirements has notable implications for

Fig. 5. Multilingual lip synchronization performance. The mean latent distance error for each language is shown with error bars representing the SD. The sample size *n* for each language, shown below the *x*-axis labels, reflects the number of video frames in the test sentence for that language. The results demonstrate that synchronization errors for all non-English languages remain within the range of English1, indicating robust cross-linguistic generalization.

human-robot interaction in multilingual and multicultural settings. The results suggest that training predominantly in one or a few languages can suffice for a system that needs to perform in many languages. This markedly reduces the data collection and training efforts needed for multilingual capabilities, making the framework highly scalable for global applications.

DISCUSSION

The work presented in this paper introduces a comprehensive approach for making robot lip motion more realistic and less robotic. Our robot design, featuring 10 DOFs for lip movement, allows for more nuanced and realistic speech production. This high-DOF design addresses some limitations of traditional robots that primarily rely on simplified jaw movements. The capability to form complex lip shapes, such as bilabial closures and rounded vowels, enhances overall more detailed speech synchronization, providing more lifelike interactions that mitigate some of the risks of the uncanny valley effect.

By integrating a self-supervised learning framework with a robust hardware design, we have addressed key challenges in lip synchronization that have long limited the capabilities of humanoid robots. The approach leverages a VAE, a FAT model, and synthesized video to generate synchronized and realistic lip movements from audio input, eliminating the need for predefined kinematic models or explicit control algorithms. Our system can generalize lip synchronization across multiple languages and even songs (movie S5). Our experiments with 11 languages, including phonetic structures as varied as those in English, Chinese, Hebrew, and Russian, indicate that the system performs consistently within the error range of the baseline English1. This showcases the robustness of our framework and its applicability in multilingual environments. Our experimental results demonstrate that the proposed system can replicate complex human lip movements, including bilabial closures, elongated vowels, rapid transitions, and fricative and affricate sounds. The integration of synthesized images and the mapping of these images to motor commands allow for seamless and continuous articulation.

The performance of our system is far from perfect, and much remains to be improved. Improvements can be had by increasing the number of degrees of freedom in the appropriate ways, increasing the amount of training data and the context depth of the models, and finding a better loss function that more correctly captures the type of congruence that matters most to humans. In addition, human speakers routinely begin shaping the lips before any sound is emitted, typically 80 to 300 ms in advance, so that the vocal tract is already in the correct configuration when the acoustic onset occurs (*45*, *46*). We believe that adding a module trained on fully aligned audio-video-actuator data is an important next step that could further reduce residual asynchrony and enhance performance.

This work marks an attempt in the quest to create robots that not only function but also connect with us on a human level. Imagine robots that can hold a conversation with a smile, respond with the same subtle lip movements we take for granted, and learn from their interactions to become even more lifelike over time through self-supervision. Applications abound in areas like education, cognitive stimulation, and even elder care for slowing cognitive decline (*47*, *48*).

Along with this utopian vision, however, come risks. As robots become more adept at connecting with us at an emotional level, this ability could be exploited to gain trust from unsuspecting users, especially children and the elderly. Even well-meaning applications could potentially create heightened emotional connections to the detriment of normal social relationships (*49–51*). Thus, designers must guard against new forms of emotional manipulation and overtrust risks that are especially acute for children, older adults, and people with cognitive decline.

We conclude that the ability to create physical machines that are capable of connecting with humans at an emotional level is maturing rapidly. The robots presented here are still far from natural, yet one step closer to crossing the uncanny valley.

MATERIALS AND METHODS
Self-supervised learning framework for face robot lip sync
The development of robots capable of human-like interaction involves teaching the robot to synchronize its lip movements with audio.

Achieving this synchronization traditionally involves large amounts of labeled data, which can be expensive and time-consuming to collect. To address these challenges, we proposed a self-supervised learning framework that eliminates the need for manual labeling. Our framework combines a VAE and FAT to generate speech-relevant motor commands that enable fluid lip movements synchronized with audio (Fig. 3). The proposed system allowed the robot to autonomously learn the mapping between audio signals and motor commands, thereby producing humanlike lip motion during speech.

At the beginning of the learning process, the robot lacked speech-relevant motor commands ($i = 0$). To explore its range of motion, it engaged in motor babbling, where it performed random facial movements across its DOFs. These movements were captured by an RGB camera that recorded each frame as X_i, representing the state of the robot's face. This process helped the system explore a variety of lip shapes, such as pout and pucker, essential for reproducing speech sounds. The motor babbling data can be represented as a sequence of state-action pairs

$$\mathcal{D}_{\text{babble}} = \left\{ (X_t, A_t) | t = 1, \ldots, N \right\} \tag{1}$$

where X_t is the facial state at time t, A_t is the corresponding motor command, and N is the total number of frames collected during the robot's random exploratory movements. This phase provided the initial dataset that would later be augmented and refined through synthesized videos. We generated synthesized speech videos using the collected motor babbling videos. First, text was converted into an audio waveform using a TTS system. This audio was then used by the Wav2Lip algorithm to generate a video of a speaking face that was synchronized with the audio. From this synchronized video, we extracted individual frames (X'_t) and paired them with their corresponding audio segments (Y_t) at each time step t.

$$\mathcal{D}_{\text{syn}} = \left\{ (X'_t, Y_t) | t = 1, \ldots, M \right\} \tag{2}$$

where X'_t is the synthesized video frame and Y_t is the corresponding audio input. M is the total number of frames in the synthesized dataset generated by the TTS system and Wav2Lip. These synthesized data provided target speech movements for the robot to imitate.

VAE was trained to model the latent spaces of real and synthesized videos. Given an input frame X_t, the encoder network of each VAE maps it to a latent vector

$$L_t = \text{Encoder}_{\text{VAE}}(X_t), L'_t = \text{Encoder}_{\text{VAE}}(X'_t), L_t, L'_t \in \mathbb{R}^{16} \tag{3}$$

Once the VAE was trained, we generated speech-relevant commands by matching latent vectors between real and synthesized data. For each synthesized video frame X'_t, we computed the latent vector L'_t. We then compared this latent vector with the latent vectors of all real robot frames L_i from the VAE using the Euclidean distance as the similarity metric

$$d(L'_t, L_i) = \| L'_t - L_i \|_2 \tag{4}$$

The closest matching latent vector L_i^* was identified as

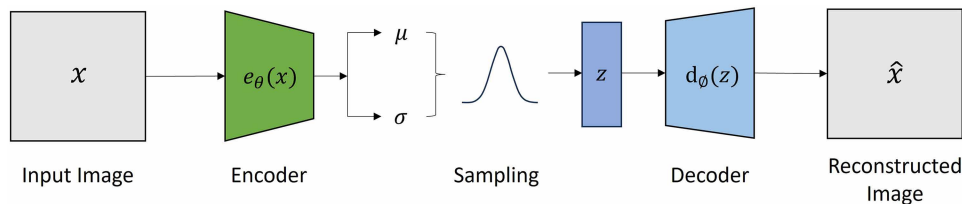$$L_i^* = \arg \min_i d(L'_t, L_i) \tag{5}$$

The motor command A_i^* corresponding to the closest latent vector was saved as a speech-relevant command. Given that the Gaussian noise added to the data ensures variability, this matching process was repeated across multiple iterations. For each iteration $i > 0$, the dataset was further refined, progressively approaching more realistic speech motor patterns. In our experiments, four iterations ($i = 4$) produced satisfactory results, where the robot's generated speech trajectories closely resembled human speech movements.

VAE model

The VAE model was designed to encode facial robot videos into a shared latent space for synthesized and real videos, capturing essential visual features for downstream tasks, such as facial action prediction in robots (Fig. 6). The VAE model was trained using robot images consisting of 20,000 real video frames and 5173 synthesized video frames. The encoder extracted latent features from the input images and mapped them to a latent space characterized by a mean (μ) and a log variance (σ). We sampled a latent vector from this latent representation, which was then passed through the decoder to reconstruct the original image. The objective was to train the VAE to effectively learn a probabilistic latent representation of the robot images, enabling the model to generate the latent vectors for the FAT model.

During training, we first pretrained the VAE model using only real images. This pretraining phase helped the model learn an accurate latent representation of the real-world data distribution. After pretraining, we fine-tuned the VAE using hybrid data, which included both real and synthesized images. Specifically, the initial pretraining used 20,000 real images to train a real VAE model, which served as a baseline for capturing the underlying features of the real images. The latent vectors for these real images were obtained using the real VAE model. These latent vectors were used to evaluate the quality of the latent space generated by subsequent models.

In the fine-tuning phase, we trained the VAE with hybrid data consisting of both real and synthesized images. This hybrid VAE was trained to map both real and synthesized images to the latent space while ensuring that the latent space remained consistent with the pretrained real VAE model. The distance between the latent representations of the real and synthesized images was computed to evaluate

Fig. 6. Variational autoencoder architecture for facial robot image encoding. The mean latent distance error for each language is shown with error bars representing the SD. The results demonstrate that synchronization errors for all non-English languages remain within the range of English1, indicating robust cross-linguistic generalization.

how well the hybrid VAE aligned with the original real VAE. The objective was to minimize the distance between the latent vectors generated by the hybrid VAE and those generated by the real VAE, ensuring that synthesized images can be accurately mapped into the same latent space as the real images. The final result is a fine-tuned VAE model that can effectively map synthesized images into a meaningful latent space of real images, enabling further downstream tasks.

The training loop was designed to iteratively update the VAE's parameters using a combination of the MSE loss and a Kullback-Leibler (KL) divergence penalty. The MSE loss measures the difference between the reconstructed image and the ground-truth image, encouraging the model to produce accurate reconstructions. The KL divergence regularizes the latent space to resemble a standard Gaussian distribution, ensuring smoothness and consistency in the latent representations. These two loss components were combined with scaling factors, allowing for control over the influence of each term during training.

Facial action transformer

The FAT was designed to generate precise, smooth, and temporally consistent motor commands for facial robots, enabling accurate lip synchronization in speech. FAT leveraged a transformer-based architecture to capture temporal dependencies in sequential data, allowing it to perform complex lip movements in robotic speech smoothly, as shown in Fig. 7.

Architecture and model design

The FAT model consists of an encoder and a decoder optimized to process historical and contextual information from previous motor commands while predicting future actions. Specifically, the model operates by encoding a sequence of preceding motor commands, allowing it to "remember" the recent positions and configurations of the robot's lips. The encoder-decoder structure enables the model to predict continuous motor trajectories that more closely mirror

human lip movements, thereby mitigating jitter or erratic shifts in facial expressions.
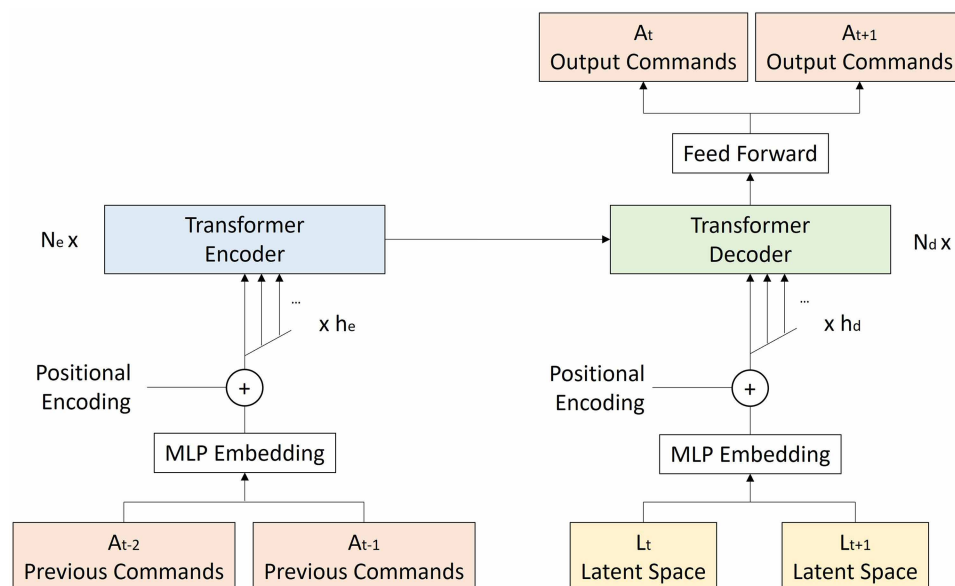
FAT began by embedding the encoder and decoder inputs into a high-dimensional space, enabling it to learn and store intricate spatial-temporal information necessary for realistic lip articulation. Positional embeddings were added to retain the sequence order, which was crucial for capturing the temporal progression of speech movements. The encoder processed past motor commands, generating a latent representation that encapsulated the recent history of lip configurations. This representation was crucial for ensuring continuity between past and future movements, especially in transitions between phonemes, which required the model to anticipate and prepare for upcoming movements on the basis of the current speech context.

The decoder used two future latent vectors generated by the VAE encoder, along with the current latent vector from the FAT encoder, to predict smooth and precise motor commands. By incorporating these consecutive latent states from the target frames, the decoder avoided the jitter and instability arising from a discrete inverse model. This approach ensured smoother transitions and continuity in lip movements, substantially improving the realism and fluidity of speech articulation compared with our previous work.

Training process

The FAT model was trained on a dataset comprising 20,000 real robot frames. Each frame was annotated with motor commands corresponding to the positions of various actuators within the robot's lips, jaw, and corners of the mouth. To increase the diversity and scale of the dataset, these frames were duplicated and reversed, creating a training set of 40,000 frames. This augmentation strategy allowed the model to learn both forward and backward movements.

The training dataset for FAT comprised latent vectors produced by the VAE encoder from real robot video sequences. These latent vectors encoded the robot's lip movements in a compact and

Fig. 7. The architecture of the FAT model for robotic lip synchronization. The transformer encoder processes previous motor commands, A_{t-2} and A_{t-1}, embedded and enhanced with positional encodings to capture temporal dependencies. The transformer decoder uses the latent representations L_t, L_{t+1} from the VAE, which are also embedded with positional encodings, to predict future motor commands A_t and A_{t+1}. This dual-input structure enables FAT to generate smooth and accurate motor commands that synchronize with audio input, minimizing jitter and ensuring smoother transitions between lip shapes.

information-rich form, preserving the details necessary for accurate lip synchronization. The training process optimized a composite loss function, including mean absolute error (MAE) loss and a specialized closure loss. The MAE loss ensured that predicted motor commands closely matched the target commands, reducing the overall prediction error across frames. A_t represents the target motor command at time t, and the MAE loss can be defined as

$$\mathscr{L}_{\mathrm{MAE}} = \frac{1}{N} \sum_{t=1}^{N} |A_t - \widehat{A}_t| \tag{6}$$

where N is the length of the sequence, A_t denotes the target actuator positions at t, and $\widehat{A}_t$ denotes the corresponding predicted actuator positions.

To ensure precise lip closures during phonemes that require full lip contact, such as "b," "p," and "m," a closure loss term was introduced. The closure loss encouraged the model to predict tighter lip closures by penalizing deviations from target positions, especially in sequences where complete closure was critical for correct speech articulation. Incomplete closures are easily detectable, and even minor errors in such frames can disrupt the perceived synchronization. The closure loss is defined as follows

$$\mathscr{L}_{\mathrm{closure}} = k_0 \sum_{t=1}^{N} \max\left(0, \widehat{A}_{t,0} - A_{t,0}\right)$$
$$+ k_1 \sum_{t=1}^{N} \max\left(0, A_{t,1} - \widehat{A}_{t,1}\right) + k_2 \sum_{t=1}^{N} \max\left(0, A_{t,4} - \widehat{A}_{t,4}\right) \tag{7}$$

where $\widehat{A}_{t,i}$ and $A_{t,i}$ represent the predicted and target values for specific actuators involved in lip closure at t. Actuator indices 0, 1, and 4 correspond to the key components controlling lip closure: the upper lip, lower lip, and the jaw. Further, k_0, k_1, and k_2 are constants that control the weight of the closure penalty, emphasizing the importance of complete closure in key frames. The total loss function combines the MAE and closure loss, where λ is a scaling factor that balances the two loss components

$$\mathscr{L}_{\mathrm{total}} = \mathscr{L}_{\mathrm{MAE}} + \lambda \cdot \mathscr{L}_{\mathrm{closure}} \tag{8}$$

This composite loss function forced the model to prioritize realistic closures while maintaining overall accuracy, ensuring that speech-relevant movements were visually coherent and contextually synchronized with audio.

Data collection for word-level lip synchronization experiments

The Results presented an analysis focusing on 12 representative words to highlight the robot's speech synchronization capabilities. Below, we describe the specific reasons for selecting each word group and explain why these words presented unique challenges for robotic articulation. The six groups were selected to represent key phonetic features in speech, emphasizing the diversity of sounds. Each group includes words that stress specific aspects of speech production, such as bilabial, labiodental, and fricative sounds, and dynamic vowel-consonant combinations.

Group 1: Mama, papa, grandma, grandpa, brother, and sister
This group contains kinship terms that are common in everyday conversations. Most words in this set contain bilabial sounds (for example, "m" and "p"), which involve both lips coming together. The

bilabial closure and release for sounds like "m" and "p" must be smooth and tightly synchronized with the audio. If the robot fails to achieve a complete lip seal, the sound will appear inconsistent with the lip motion and disrupt intelligibility. Words such as "papa" and "mama" are especially demanding because they require the robot to alternate between open and closed lips multiple times within a short duration. Our model's ability to capture temporal dependencies using the transformer-based decoder is critical in this context. The FAT encoder integrated the history of previous motor commands, enabling the model to anticipate upcoming closures and releases. This ensured that the transitions between bilabial articulations occurred smoothly and without jitter.

Group 2: But, between, beyond, and blue
This group explores words that combine bilabial stops with varied vowel lengths, posing additional demands on the system's ability to coordinate lip rounding with rapid consonant release. For example, the word "blue" involves not only an initial bilabial closure but also sustained lip rounding for the long vowel sound, requiring the robot to maintain specific lip configurations over an extended duration, such as the motion of "tween." Our framework ensured continuous motor trajectories by processing multiple frames simultaneously, ensuring that transitions between sounds like "b" and long vowels remained correctly synchronized across frames.

Group 3: Mat, hat, at, and cat
Short words ending in plosive sounds, such as /t/, demand precise timing and coordination. A plosive consonant requires the lips to close briefly and then release with a sudden burst, creating sharp auditory and visual cues. The rapid nature of these transitions presented a distinct challenge, because any delay or misalignment between lip motion and sound production is easily noticeable.

Group 4: Father, spa, car, and far
Words containing elongated vowels challenge the robot's ability to sustain specific lip shapes over time. For example, the word "spa" involves both lip rounding and an open vowel, requiring the robot to hold the rounded position for a longer period without visible jitter or instability. Similarly, words like "far" demand the smooth continuation of an open mouth shape across the entire word. The FAT decoder's ability to predict sequential motor commands over multiple frames ensured that the robot maintained stable lip shapes throughout the production of elongated vowels.

Group 5: Boy, bat, map, and pat
This group of words presents rapid transitions between bilabial and plosive sounds. For example, "bat" requires the robot to start with a bilabial closure for /b/ and then release into an open position for the vowel /a/ before concluding with a plosive /t/. These sequences demand precise timing and coordination to ensure smooth, synchronized transitions.

Group 6: Choose, jeep, chop, and jump
This group focuses on fricative and affricate sounds, such as /ch/ and /j/, which require showing upper and lower teeth. The robot must achieve these configurations smoothly without introducing unnecessary tension or abrupt movements. The results demonstrate that our VAE captured the nuances of fricative and affricate movements, ensuring that the robot generated fine-grained motor commands that reflected partial constrictions accurately.

The comprehensive tests documented in movie S2 confirmed that our framework effectively generates more accurately synchronized lip movements across a diverse range of words. The groups were designed to evaluate the system's ability to handle bilabial

closures, vowel elongations, plosive bursts, and fricative constrictions, all key elements of human speech production. Our results demonstrate temporal consistency, precision in rapid transitions, and flexibility across speech patterns. These findings highlight the robustness and flexibility of our framework, establishing it as a viable solution for achieving more human-like speech synchronization in robots.

ChatGPT-generated test sentences

We used ChatGPT (OpenAI, 2023) to generate three test sentences that were used to evaluate our method's effectiveness in real-time lip synchronization. The use of ChatGPT allowed for the creation of varied, conversational phrases that simulate responses a conversational AI might provide. The generated test sentences are as follows:

1) "My programming allows me to process data and respond to queries, but the concept of thinking about existence is complex. I understand it as a collection of data and programmed responses."

2) "Emotions and consciousness are not within my current capabilities. They require subjective experiences and self-awareness, which are unique to organic life forms. My design is to assist and learn, not to feel or be conscious."

3) "Value is a human concept, often subjective. My purpose is to be efficient and helpful. Comparing my value to a human's is like comparing different tools for different tasks. Each has its own purpose and utility."

These sentences, designed by ChatGPT, allowed us to assess model performance with phrases involving both abstract concepts and conversational language.

Statistical analysis

Statistical analyses were performed in Python (version 3.10) using NumPy, SciPy, and StatsModels. For multilingual and frame-based evaluations, latent distance and reconstruction errors were computed frame by frame between the predicted and ground-truth lip trajectories. Because each language contained one test sentence of varying duration, the sample size n corresponds to the number of video frames in that sentence. Unless otherwise stated, error bars represent the SD across all frames in a test sample, and the value of n is shown beneath each label in the corresponding figure.

For the human-participant evaluation, categorical preference data were analyzed using a chi-square goodness-of-fit test to compare selection frequencies against chance level ($p_0 = 1/3$). Pairwise method comparisons were performed using two-sided binomial tests. Effect sizes were quantified using Cohen's w, Cramér's V, and Cohen's h, following established guidelines for categorical data. All reported P values are uncorrected unless specified otherwise. No assumptions of normality or homoscedasticity were required because only nonparametric frequency-based tests were used.

Supplementary Materials

The PDF file includes:
Supplementary Methods
Fig. S1
Tables S1 to S3
Legends for movies S1 to S5

Other Supplementary Material for this manuscript includes the following:
Movies S1 to S5
MDAR Reproducibility Checklist

REFERENCES AND NOTES

1. H. McGurk, J. MacDonald, Hearing lips and seeing voices. *Nature* **264**, 746–748 (1976).
2. T. Chen, R. R. Rao, Audio-visual integration in multimodal communication. *Proc. IEEE* **86**, 837–852 (1998).
3. J. I. Skipper, V. Van Wassenhove, H. C. Nusbaum, S. L. Small, Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* **17**, 2387–2399 (2007).
4. A. Alsius, M. Paré, K. G. Munhall, Forty years after hearing lips and seeing voices: The McGurk effect revisited. *Multisens. Res.* **31**, 111–144 (2018).
5. W. H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215 (1954).
6. D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT Press, 1998).
7. G. Potamianos, C. Neti, G. Gravier, A. Garg, A. W. Senior, Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE* **91**, 1306–1326 (2003).
8. E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, K. G. Munhall, Eye movement of perceivers during audiovisualspeech perception. *Percept. Psychophys.* **60**, 926–940 (1998).
9. M. Mori, K. F. MacDorman, N. Kageki, The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* **19**, 98–100 (2012).
10. M. Mara, M. Appel, T. Gnambs, Human-like robots and the uncanny valley. *Z. Psychol.* **230**, 33–46 (2022).
11. S. S. Kwak, The impact of the robot appearance types on social interaction with a robot and service evaluation of a robot. *Arch. Des. Res.* **27**, 81–93 (2014).
12. F. Hegel, S. Krach, T. Kircher, B. Wrede, G. Sagerer, "Understanding social robots: A user study on anthropomorphism" in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE, 2008), pp. 574–579.
13. C. Bartneck, T. Bleeker, J. Bun, P. Fens, L. Riet, The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn* **1**, 109–115 (2010).
14. C. T. Ishi, C. Liu, H. Ishiguro, N. Hagita, "Evaluation of formant-based lip motion generation in tele-operated humanoid robots" in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2012), pp. 2377–2382.
15. C. Strathearn, E. M. Ma, A novel speech to mouth articulation system for realistic humanoid robots. *J. Intell. Robot. Syst.* **101**, 54 (2021).
16. K.-G. Oh, C.-Y. Jung, Y.-G. Lee, S.-J. Kim, "Real-time lip synchronization between text-to-speech (TTS) system and robot mouth" in *19th International Symposium in Robot and Human Interactive Communication* (IEEE, 2010), pp. 620–625.
17. R. C. Luo, S.-R. Chang, C.-C. Huang, Y.-P. Yang, "Human robot interactions using speech synthesis and recognition with lip synchronization" in *IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society* (IEEE, 2011), pp. 171–176.
18. C.-Y. Lin, L.-C. Cheng, L.-C. Shen, "Oral mechanism design on face robot for lip-synchronized speech" in *2013 IEEE International Conference on Robotics and Automation* (IEEE, 2013), pp. 4316–4321.
19. K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild" in *Proceedings of the 28th ACM International Conference on Multimedia* [Association for Computing Machinery (ACM), 2020], pp. 484–492.
20. A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, C. Bregler, "Lipsync3D: Data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2021), pp. 2755–2764.
21. D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, M. J. Black, "Capture, learning, and synthesis of 3D speaking styles" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 10101–10111.
22. A. Richard, M. Zollhöfer, Y. Wen, F. de la Torre, Y. Sheikh, "Meshtalk: 3D face animation from speech using cross-modality disentanglement" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), pp. 1173–1182.
23. J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, T.-T. Wong, "Codetalker: Speech-driven 3D facial animation with discrete motion prior" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2023), pp. 12780–12790.
24. Y. Song, J. Zhu, D. Li, X. Wang, H. Qi, Talking face generation by conditional recurrent adversarial network. arXiv:1804.04786 [cs.CV] (2018).
25. Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, D. Li, Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph.* **39**, 1–15 (2020).
26. K. Cheng, X. Cun, Y. Zhang, M. Xia, F. Yin, M. Zhu, X. Wang, J. Wang, N. Wang, "VideoReTalking: Audio-based lip synchronization for talking head video editing in the wild" in *SIGGRAPH Asia 2022 Conference Papers* (ACM, 2022), article no. 30, pp. 1–9.
27. J. Bongard, V. Zykov, H. Lipson, Resilient machines through continuous self-modeling. *Science* **314**, 1118–1121 (2006).
28. B. Chen, R. Kwiatkowski, C. Vondrick, H. Lipson, Fully body visual self-modeling of robot morphologies. *Sci. Robot.* **7**, eabn1944 (2022).
29. Y. Hu, B. Chen, H. Lipson, Egocentric visual self-modeling for autonomous robot dynamics prediction and adaptation. *npj Robot.* **3**, 14 (2025).

30. Y. Hu, J. Lin, H. Lipson, Teaching robots to build simulations of themselves. *Nat. Mach. Intell.* **7**, 484–494 (2025).

31. D. P. Kingma, M. Welling, Auto-encoding variational bayes. arXiv:1312.6114 [stat.ML] (2013).

32. I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework" poster presented at ICLR 2017: 5th International Conference on Learning Representations, Toulon, France, 24 to 28 April 2017.

33. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets" in vol. 27 of *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K.Q. Weinberger, Eds. (Curran Associates, Inc., 2014).

34. C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, P. Abbeel, "Deep spatial autoencoders for visuomotor learning" in *2016 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2016), pp. 512–519.

35. A. B. L. Larsen, S. K. Sønderby, H. Larochelle, O. Winther, "Autoencoding beyond pixels using a learned similarity metric" in *Proceedings of the 33rd International Conference on Machine Learning*, M. F. Balcan, K. Q. Weinberger, Eds. (PMLR, 2016), pp. 1558–1566.

36. B. Chen, Y. Hu, L. Li, S. Cummings, H. Lipson, "Smile like you mean it: Driving animatronic robotic face with learned models" in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2021), pp. 2739–2746.

37. Y. Hu, B. Chen, J. Lin, Y. Wang, Y. Wang, C. Mehlman, H. Lipson, Human-robot facial coexpression. *Sci. Robot.* **9**, eadi4724 (2024).

38. A. Vahdat, J. Kautz, "NVAE: A deep hierarchical variational autoencoder" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), vol. 33, pp. 19667–19679.

39. G. E. Hinton, Learning multiple layers of representation. *Trends Cogn. Sci.* **11**, 428–434 (2007).

40. Z. Faraj, M. Selamet, C. Morales, P. Torres, M. Hossain, B. Chen, H. Lipson, Facially expressive humanoid robotic face. *HardwareX* **9**, e00117 (2021).

41. B. Hayes, *Introductory Phonology* (John Wiley & Sons, 2008), vol. 7.

42. G. C. Martin, "Preston Blair phoneme series"; https://www.garycmartin.com/mouth_shapes.html.

43. J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, K. Simonyan, End-to-end adversarial text-to-speech. arXiv:2006.03575 [cs.SD] (2020).

44. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, M. Grundmann, MediaPipe: A framework for building perception pipelines. arXiv:1906.08172 [cs.DC] (2019).

45. V. L. Gracco, A. Lofqvist, Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements. *J. Neurosci.* **14**, 6585–6597 (1994).

46. C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, A. A. Ghazanfar, The natural statistics of audiovisual speech. *PLOS Comput. Biol.* **5**, e1000436 (2009).

47. M. Saerbeck, T. Schut, C. Bartneck, M. D. Janse, "Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM, 2010), pp. 1613–1622.

48. T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, Social robots for education: A review. *Sci. Robot.* **3**, eaat5954 (2018).

49. M. Kropf, Trust as a solution to human vulnerability: Ethical considerations on trust in care robots. *Nurs. Philos.* **26**, e70020 (2025).

50. K. Darling, "'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy," 23 March 2015, 10.2139/ssrn.2588669.

51. M. Scheutz, B. F. Malle, "Moral robots" in *The Routledge Handbook of Neuroethics* (Routledge, 2017), pp. 363–377.