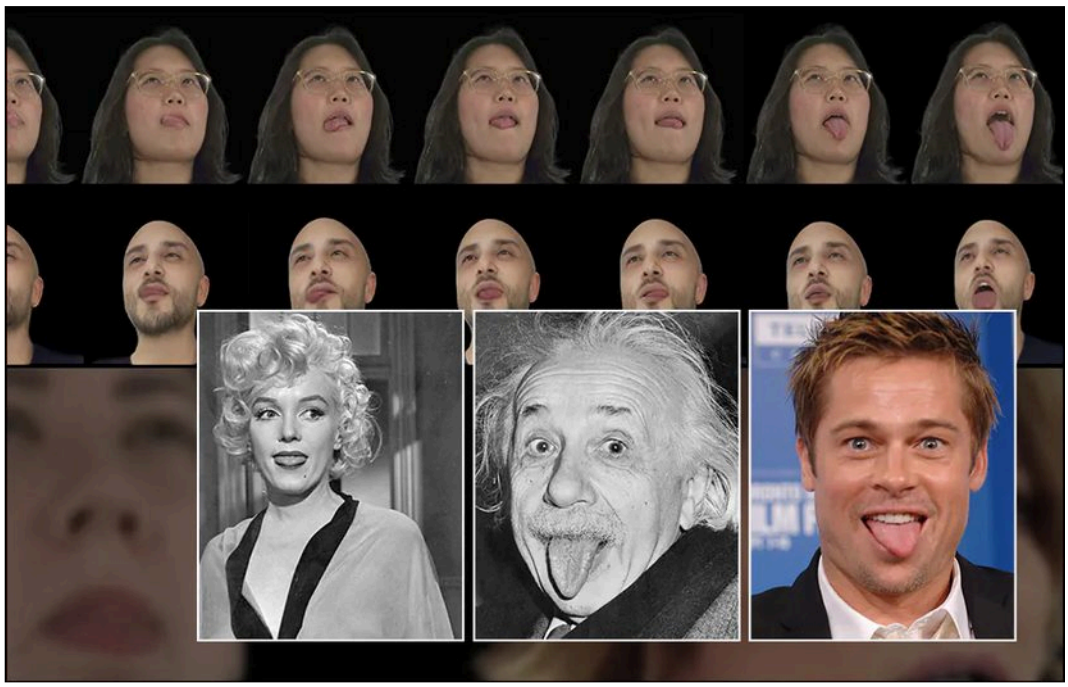
No Tongues! AI Researchers Address Missing Anatomy in Facial Animation

Originally published Weds 11 Nov 2026 at https://www.unite.ai/no-tongues-ai-researchers-address-missing-anatomy-in-facial-animation/



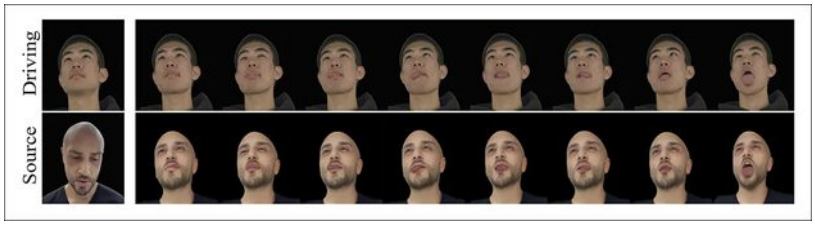
Missed in 2026: *It's that time of year when the research scene dips and spikes in quality and volume in lockstep with the last few conferences of November and December. To traverse these sporadic famines, we're taking a look at some of the more interesting papers that nearly made the cut in 2026, but were ousted by other works, in more competitive submission periods.*

Until recently, deepfake practitioners, amateur and professional alike, had bigger concerns to address than whether or not the reconstructed human appeared to have a tongue. After all, the tongue is the most mercurial of facial features – not evident at all in 'mumblers', and only intermittently observed in most normal contexts.

Consequently, when you do need it, as far as AI reconstructions are concerned, it's not there. In a climate where amateur diffusion-based reconstruction has reached a stunning level of realism, and our critical faculties in regard to AI humans is becoming ever more discerning, it is perhaps time to address the issue.

Licking the Problem

In July an academic collaboration between Australia and China offered a pipeline to address the lack of tongue simulation in AI workflows. Titled *TongueReenact*, the proposed framework throws practically every method in the deepfake toolkit at the problem, including Gaussian Splatting, parametric humans, Neural Radiance Fields (NeRF), and diffusion-based video generation (sorry GANs, you didn't make this cut).



Above, the driver (real) capture whose characteristics are transferred onto the TongueReenact render underneath. Source

Though the project's promised code does not appear to have been delivered since the paper was published, and though no supplementary video is yet released, this is nonetheless a landmark offering, since it is the first to address the tongue in AI-based reconstruction, outside of the medical sphere – and to address the critical lack of annotated data, which is essential for any proposed solution.

The researchers made innovative use of diverse existing frameworks and data, in order to create a tongue-segmentation framework based on SegmentAnything, and to generate pseudo-labels for the newly-segmented data.

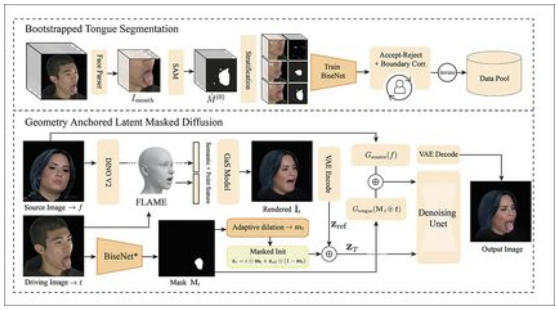
The paper states:

> *'[We] present the first framework for cross-identity tongue dynamics transfer in face reenactment. Given a source frame and a driving video, our method transfers both facial motion and tongue articulation onto the source identity.*

© Martin Anderson and/or commissioning entities

> *'The central insight of our work is that tongue synthesis cannot be treated as a generic inpainting problem. The tongue must be spatially grounded to the source face geometry, which itself changes with every driving frame.'*

The system begins by transferring the driving person's facial movements onto the source face, using 3D Gaussian Splatting and the FLAME facial model. This produces an animated face with the correct pose and expression, but still lacking a tongue.



The TongueReenact workflow. Top: Tongue images are identified, segmented and manually refined to train a detection model. Bottom: Facial movements are transferred from a driving video to a source face, before a diffusion model generates the missing tongue in the appropriate position.

A Stable Diffusion model is then used to generate the missing tongue, guided by its appearance in the driving video and the geometry of the reconstructed face.

Regarding the need to address this shortfall in facial reconstruction, the authors observe*:

> *'During expressive actions such as laughter, disgust, or surprise, [the tongue] can dominate the visible oral region entirely. Yet nearly all existing reenactment and avatar systems either model the oral interior as a textureless void or leave it entirely unconstrained.*

> *'This neglect is not coincidental, as the tongue is a uniquely transient structure, fully concealed at rest, intermittently exposed during speech, and suddenly dominant in extreme expressions, making it fundamentally resistant to the static parametric modeling applied to the rest of the face.*

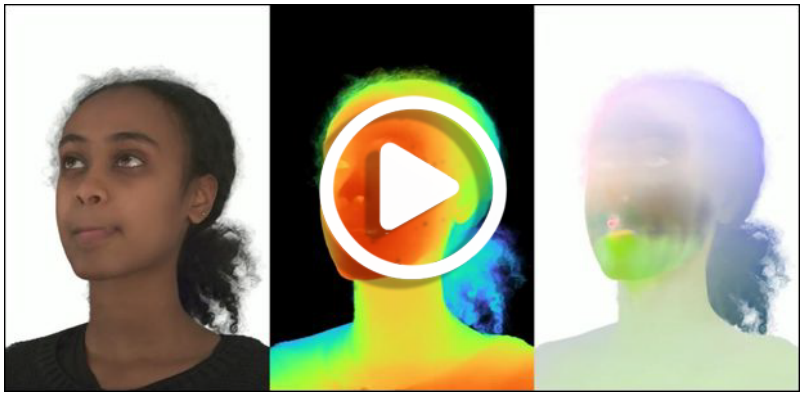
> *'Addressing this overlooked challenge is the central motivation of our work.'*

In tests, the new method substantially outperformed five competing facial animation systems, achieving a more than 100% accuracy advantage over its closest rival on key measures of tongue positioning and size – though it has to be conceded that tests of this kind are hard to contrive when one is breaking new ground, and when prior frameworks address the new challenge minimally, or not at all.

Let's take a closer look at the new approach.

Method

To develop a system capable of identifying tongues in ordinary facial footage, the NERSemble dataset of multi-view recordings of human heads was used alongside synthetically generated frames from the VFHQ dataset, which features over 16,000 high-quality facial videos:

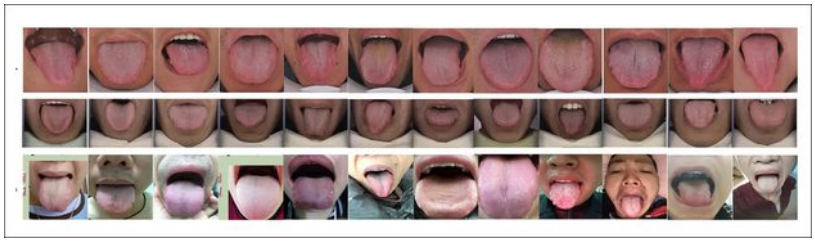


CLICK TO PLAY VIDEO ON SITE

Press to play *An interestingly 'lingual' example of Radiance Field reconstruction from the Nersemble project site. However, the tongue seen here is treated as generic geometry, and has no distinct domain.* Source

The MaskGAN face-parsing model was used to locate the mouth in each image, after which the TongueSAM medical detector was used to identify the likely tongue region, before passing it to SegmentAnything, from which an initial outline, (i.e., pseudo-label) of the visible tongue was generated.

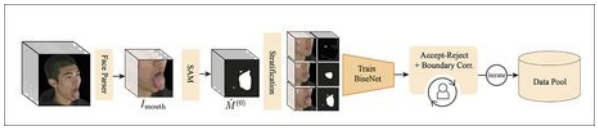
© Martin Anderson and/or commissioning entities



Samples from the TongueSAM dataset, intended for analytical medical usage rather than evaluation of facial expressions. Source

The resulting masks were divided into three groups, according to how much of the tongue was visible. Incorrect examples were discarded, and promising masks manually corrected to distinguish the tongue from surrounding lips, teeth, and other mouth tissue. These refined annotations were used to train the BiSeNet image segmentation network.

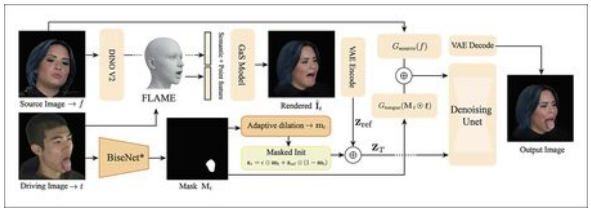
The model was then applied to additional NERSemble subjects and synthetic VFHQ frames, with its predictions checked and corrected by humans before further training – a process that was repeated until performance on previously unseen subjects stabilized:



Pipeline for parsing existing data into 'tongue-domain' data.

For tongue synthesis, the oral region identified by BiSeNet was slightly enlarged before being passed to the diffusion model, giving the generated tongue room to extend beyond the initial segmentation, as the mouth changed shape.

Generation was performed in the latent space of Stable Diffusion Image Variations, with separate guidance derived from the reconstructed facial geometry and the tongue visible in the driving frame, and only the identified mouth/tongue region regenerated.



The generative pipeline. The DINO V2 reference in upper left quadrant encapsulates the use of GAGAvatar (see below), which use DINO V2 appearance features in concert with FLAME.

Data and Tests

BiSeNet was trained on two output classes derived from the aforementioned data processing stages: *background* and *tongue*. Training was carried out over 50,000 iterations at a learning rate of 10^{-3}, using 512x512px images on batch size 4, on a single NVIDIA A40 GPU with 48GB of VRAM. The training data consisted of the aforementioned NERSemble and VFHQ frames.

For the reenactment portion, the pretrained GAGAvatar Gaussian Splatting backbone was used. The Stable Diffusion Image Variations checkpoint was also used, with the VAE and CLIP encoder frozen. The Guidance encoders and denoising UNET were then trained on the NERSemble data for 85,000 iterations at batch size 8, with 512×512 inputs. For inference, 20 DDIM steps were used (for many current diffusion models, this is considered a high amount, designed to favor detail, with certain inevitable tradeoffs).

Rival avatar frameworks tested were GPAvatar; the feted LivePortrait; X-NeMo; and X-Portrait. Each of these was evaluated against 9,000 cross-identity VFHQ source/driving pairs.

Metrics used were Learned Perceptual Image Patch Similarity (LPIPS); *Tongue Presence* (the proportion of frames where a tongue is detected); *Tongue IoU* (the overlap between the generated and driving tongue regions); and *Tongue Area Similarity* (how closely the generated tongue's size matched the driving tongue).

Metric	Ours	X-NeMo	GPAvatar	LivePortrait	Portrait4D	X-Portrait
LPIPS (tongue) ↓	**0.2306**	0.3937	0.2890	0.3454	0.2591	0.4209
Tongue Presence ↑	**0.7891**	0.7367	0.4274	0.5986	0.5102	0.5067
Tongue IoU ↑	**0.3582**	0.1400	0.1387	0.1603	0.1918	0.1508
Tongue Area Similarity ↑	**0.4338**	0.1180	0.0713	0.1255	0.0818	0.1330

Test results comparing TongueReenact with five rival face-reenactment frameworks. TongueReenact achieves the best result across all four tongue-specific metrics; arrows indicate whether lower or higher scores are preferred.

Of the initial comparison, the authors state:

> *'Our method achieves the best performance across all four tongue transfer metrics, with a Tongue IoU of 0.3582 and Tongue Area Similarity of 0.4338, more than double the nearest competitor, confirming substantially more accurate tongue dynamics transfer.*
>
> *'Tongue Presence of 0.7891 further demonstrates that our method reliably synthesizes visible tongue articulation where existing methods frequently produce none.'*

© Martin Anderson and/or commissioning entities



Some highlights from some of the extensive qualitative examples shown in the source paper. Here we see comparisons between the new approach and the five pı are equipped to generate tongues. Please refer to the original paper for more examples and superior resolution.

Though it's not something that we can judge for ourselves, due to the current lack of supplementary video material, the authors also tested the system for temporal consistency using T-LPIPS:

Metric	With Tongue	Without Tongue
T-LPIPS ↓	**0.0149**	0.0291
Flow Jitter ↓	**0.2185**	0.6229
Flow Magnitude ↓	**0.5132**	1.2541

Results of tests for temporal consistency.

Here the authors comment:

> *'Our full model achieves substantially lower flow jitter (0.218543 vs. 0.622901) and flow magnitude (0.513165 vs. 1.254073), confirming that geometry-anchored denoising produces smooth and stable tongue motion.*
>
> *'Lower T-LPIPS further confirms that tongue synthesis does not compromise temporal coherence.'*

Conclusion

The uncanny valley effect is being gradually overcome by some of the latest diffusion-based video models. However, there is a certain novel psychological barrier beginning to manifest, whereby people have a negative take on AI-generated material for reasons other than the quality of what they are watching.

This is set to complicate the business of qualitative evaluation, though I suppose that metrics based on 'blind' tests (where the user wants to downgrade AI output, but fears to downgrade real output) could help to develop a kind of 'shadow' scale for AI realism.

Therefore, though there are strong research strands for the reproduction via AI of secondary identity traits such as hair, emotion simulation, gaze, and now tongue movement, they may, for the moment, be moot points.

* *My conversion of the authors' inline citations to hyperlinks.*

First published Wednesday, November 11, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Website: martinanderson.ai
Contact: martin@martinanderson.ai