

FedFAn: Federated Learning for Facial Animation

Jack Parry, Vinay Namboodiri

University of Bath, UK
jp2793@bath.ac.uk, vpn22@bath.ac.uk

Abstract

We present FedFAn: a system for training audio-driven facial animation models across organisational silos via federated learning. AI model providers often serve companies who hold sensitive biometric and proprietary data which is used for model inference and would make valuable training data, but cannot be shared with the provider. In this paper we begin by outlining the problem of privacy in facial animation, defining three key groups for whom privacy is a consideration: trainers, managers and drivers. Trainers of avatar facial animation serve managers (such as studios) that hold valuable driver performance data and character avatars, but cannot reasonably centralise those assets with the trainer. FedFAn allows collaborative training by separating the identity-agnostic model components from the components that leak identifying information: the audio encoder is trained only during public pretraining and frozen during private federated learning, speaker style embeddings remain local to each client, and secure aggregation prevents the server from observing individual updates. We show that a model pretrained on the VOCASET dataset and then privately finetuned with federated learning on the HDTF dataset improves adaptation to unseen speakers, showing that decentralised data can improve cross-domain robustness, which benefits both trainers and managers. We evaluate the privacy limits of this architecture using membership inference attack, model-internal speaker identification, gradient leakage analysis, and a differential privacy feasibility study. These experiments show where speaker information (which can identify drivers) is carried, informing our model architecture and data handling approach.

Code — <https://github.com/Jack-Paz/FedFAn>

Supplementary Material —

<https://github.com/Jack-Paz/FedFAn/blob/main/parry-fedfan-supplementary-2026.pdf>

1 Introduction

Large-scale avatar training has arrived. Recent work from Meta (Li et al. 2026) has set a new standard in avatar animation by training codec avatars on 1M videos and 2737 subjects drawing on the success of data scaling in text (Devlin et al. 2019; Brown et al. 2020; Achiam et al. 2023) and video models (Radford et al. 2021; He et al. 2022). More diverse

data can improve fine-grained animation and generalisation to unseen identities, but the scale of biometric data collection required presents serious governance concerns when data is centralised by a small number of technology providers.

We seek to address this problem as it relates to a real-world industry: small video game and animation studios licensing facial animation systems for lip-sync and expression generation.

In this work, we aim to outline the different interest-holders for such systems: the video game studios that we term *managers*, the companies that train the facial animation software we term *trainers*, and the end users as well as others such as voice actors who use the system that we term *drivers*, and present the privacy concerns for these interest-holders.

We propose a system in which a trainer of a facial animation system trains a shared model with multiple managers via federated learning (FL) (McMahan et al. 2017a): a training paradigm in which participants collaboratively train a shared model by exchanging only model updates, while the raw training data never leaves each participant’s premises. The trainer operates the central server and distributes model updates, while managers hold local datasets containing driver voice recordings, corresponding high-quality facial motion, and learned character specific style information. Specifically, studios will use a model to generate the first-pass animation for their characters from voice actor recordings. Any errors found in the outputs would be corrected by animators during quality assurance. These corrections are human-verified vertex sequences paired with the original audio, combining over multiple productions to form a local dataset of useful training data. This describes a realistic governance problem: the trainer can benefit from increased training data, but managers have legal and commercial commitments to safeguarding biometric driver data or unreleased creative assets. We are therefore presenting a technical solution to a societal problem which has implications far beyond facial animation: *who controls the infrastructure through which biometric and creative assets are converted into AI training data?*

Privacy in this setting requires protecting three things: voice identity, character styles, and training data membership. We train a prototypical audio-driven facial animation system Imitator (Thambiraja et al. 2023) in a cross-silo fed-

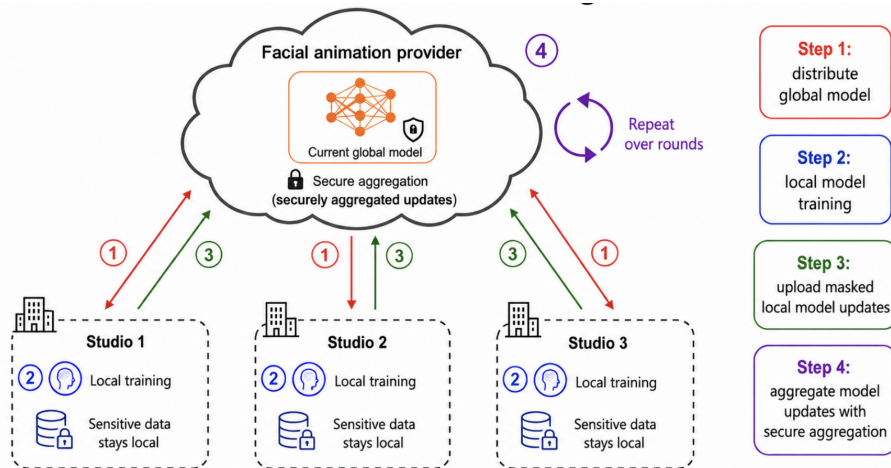


Figure 1: An overview of our cross-silo federated learning framework: For each round of training, the global model is distributed to each silo, local models are updated over multiple epochs of training on local datasets, before being securely shared and aggregated via secure aggregation.

erated learning setting with secure aggregation (Bonawitz et al. 2016). We ensure that the parts of the model that encode speaker information are frozen, and style-specific avatar information learned by the model is not shared with the central server. The central result is that privately finetuning a baseline model in a federated setting improves model generalisation to new character styles.

To evaluate the level of system privacy, we apply membership inference attacks (MIA), speaker identification at four model-internal layers and gradient leakage analysis. We find that freezing the audio encoder reduces speaker-level MIA AUC from 0.93 (near-perfect) to 0.61 (where chance= 0.5), and that without the private style embedding, avatar style identification falls to chance levels even at style-conditioned shared layers.

Our core contributions are as follows:

- FedFAn: A novel application of federated learning to audio-driven facial animation to reduce centralised control over biometric and creative assets.
- We show that centralised pre-training followed by private FL outperforms the baseline centralised model on unseen speaker adaptation.
- We propose and evaluate a privacy architecture for federated learning in facial animation: a frozen audio encoder, private style embeddings, and secure aggregation of shared component updates.
- A privacy evaluation of an audio-driven facial animation system: MIA, model-internal speaker identification, gradient leakage analysis, and a speaker-level differential privacy feasibility analysis.

2 Related Work

2.1 Audio-Driven Facial Animation

Some audio-driven facial animation approaches aim to directly generate 2D talking head videos (Xu et al. 2024a).

However, approaches which use underlying 3D representations provide greater control for downstream applications such as video game production (Edwards et al. 2020). Our federated learning approach builds on Imitator (Thambiraja et al. 2023) which uses the FLAME head model (Li et al. 2017). FLAME is a 3D morphable model (3DMM): a template face mesh whose shape and expression are controlled by a compact set of learned parameters, allowing any facial pose to be represented and animated by adjusting those parameters. An early audio-driven 3D facial animation system, VOCA (Cudeiro et al. 2019) succeeded in animating the FLAME 3DMM from audio input; FaceFormer improved upon this work by using transformer models and predicting vector-space offsets rather than FLAME expression parameters (Fan et al. 2022), and Imitator (used in this paper) extended FaceFormer with rapid style conditioning to new identities (Thambiraja et al. 2023). Later work (Xing et al. 2023) aimed to learn latent codes for animating both the FLAME and BIWI mesh topology. We choose to utilise the FLAME 3DMM because it is widely-used as an underlying controllable representation which can be built upon with techniques such as NeRF (Zielonka, Bolkart, and Thies 2023), gaussian splatting (Xu et al. 2024b) and diffusion (Ding et al. 2023) for improved texture and fine detail generation. Improving lip-sync, personalisation and expression control for the FLAME model would therefore provide a better base for state of the art models to improve upon.

Non-FLAME based facial animation approaches include (Guo et al. 2021) which trains an audio-driven NeRF avatar.

Scaling models and dataset sizes through self-supervised pretraining in text (Devlin et al. 2019; Brown et al. 2020; Achiam et al. 2023) and vision models (Radford et al. 2021; He et al. 2022), has proven powerful. Similar efforts have been made in the avatar domain, beginning with the advent of First Order Motion Models (FOMM) (Siarohin et al. 2019). Recently, Meta’s large-scale codec avatars (LCA) (Li et al. 2026) has succeeded in scaling full-body animat-

able 3D gaussians (similar to (Qiu et al. 2025)) by separately learning to output a canonical representation which can then be adapted to expressive, avatar-specific animation. This motivates our approach, training our character-agnostic viseme decoder via federated learning to improve character-specific performance after finetuning. Crucially, this related work takes data centralisation as a given for large-scale training, however, we show that it is possible for distinct silos to retain data sovereignty while improving a facial animation model.

2.2 Federated Learning

Federated learning (McMahan et al. 2017a) enables training on decentralised datasets without sharing sensitive information. We are not aware of prior work applying FL to facial animation, although FL has been applied to related fields such as facial recognition (Aggarwal, Zhou, and Jain 2021; Meng et al. 2022) and expression recognition (Shome and Kar 2021). We consider cross-silo federated learning (Huang, Huang, and Liu 2022) as the most suitable setting for collaborative model training across studios (see Figure 1).

Parameter freezing has been combined with FL before for example for improving efficiency (Chen et al. 2021). In this paper we freeze the audio encoder to avoid identification of training speakers, which has the added benefit of reducing overall trainable model size.

Secure aggregation (SA) (Bonawitz et al. 2016) is a natural extension to FL, by computing a mask which cancels out upon aggregation, client updates are obfuscated from the central server and the resulting model is identical. This has limitations in cross-device FL due to client dropout, but is quite practical for our cross-silo use-case.

2.3 Privacy in Federated Learning

Federated learning on its own does not provide any privacy guarantees (AbdulRahman et al. 2020). Despite training data never leaving the device or silo, many attacks have been developed that aim to extract information about the training data from either the trained model or communicated client updates. Gradient inversion (Geiping et al. 2020) shows you can fully reconstruct training data from parameter updates communicated to the central server. One solution to this which we also implement is secure aggregation (Bonawitz et al. 2016; Bell et al. 2020). Membership inference attacks (MIA) (Yeom et al. 2018) show that it is possible to identify if a given data item was used during training; this naturally generalises to identity-level MIA, in which the attacker aggregates evidence across all of a target individual’s data, compromising its privacy. Differential privacy (DP) (McMahan et al. 2017b; Andrew et al. 2021) is often coupled with FL to provide formal guarantees against MIA but introduces performance degradation, particularly in smaller silo-FL settings, or as model complexity increases. We present a rigorous feasibility evaluation of user-level DP in Section 6.4, finding that it is not suitable at our training scale. This motivates our decision to analyse and mitigate privacy leakage through selective layer freezing and secure aggregation.

2.4 Biometric Data and Regulation

Modern AI scaling laws reward companies that collect biometric and behavioural data at large scale. Recent work has shown that this applies to avatar identity generalisation too (Li et al. 2026). This concentrates capability among the small number of actors that can lawfully and contractually aggregate such data. Voice recordings, facial motion capture, and proprietary character styles are both sensitive biometric data, and valuable creative assets.

Regulatory frameworks acknowledge this dual nature. Facial geometry and voice recordings are classified as “special” personal data under Article 9 of the GDPR (Protection 2018), as protected biometric identifiers under the Illinois Biometric Information Privacy Act (Illinois General Assembly 2008), and as restricted-use inputs in workplace settings under the EU AI Act (EU AI Act 2024). In the games and entertainment industries, safe and legal data storage requirements are compounded by contractual requirements like voice actor agreements, character-likeness rights, and pre-release confidentiality (Edwards et al. 2020). This motivates our work; we present a system where distinct silos can collaborate to improve a shared AI model while retaining data sovereignty.

3 Institutional Setting and Threat Model

In this section, we define three key interest-holders in facial animation, discuss their respective adversaries, and the threats they represent. The three interest-holders are: “trainers”, “managers” and “drivers” (see Figure 2). Each interest-holder possesses some sensitive data they wish to keep from adversaries. In the case of drivers, the data is personally-identifying data as defined by GDPR. For trainers and managers, adversaries would be targeting industry secrets or intellectual property (IP). We aim to highlight these distinctions and explain how the present work addresses each concern individually.

3.1 Training Facial Animation Systems

Trainers of facial animation software are for the purposes of this study defined as the individual or group who train the AI models. They set the model parameters, store the training data and execute the training process.

Training state-of-the-art facial animation systems requires large amounts of sensitive video and audio data (Li et al. 2026), presenting a natural centralisation risk. Trainers must also be careful to source and store their training data in ways that comply with regulations.

Adversarial threats in this case are predominantly from hackers or rival firms who could aim to steal IP such as model weights, training data, and the training process. By allowing data to remain with managers, our system reduces this attack vector. It is also important to restrict who has access to the trained model due to ethical issues arising from the outputs of facial animation systems, as they can be used for deepfake technology (Tolosana et al. 2020), but this is out of scope for our architecture.

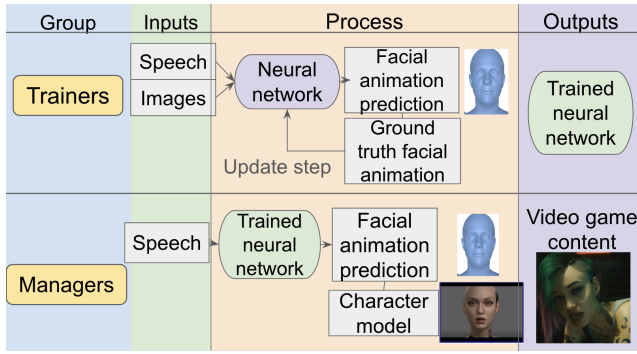


Figure 2: An illustration of the inputs, process and outputs of trainers and managers of facial animation systems. Drivers are discussed in Section 3.3.

3.2 Managing Facial Animation Systems

Managers are entities that decide where and how the output of the animation model should be used. Examples of managers are video game studios, animation studios, and social media companies. Note that managers may also be trainers of facial animation software, but they often are not, and therefore “trainers” are a separate category. The threat to managers comes from the media and individuals who wish to leak confidential information such as release schedules or story points, and rival studios who may aim to seize IP such as video game assets. Any leak can be disastrous to managers (Berchtold et al. 2013), so they may even be reluctant to share internal data with trainers. Our approach allows managers to retain sovereignty over their protected assets while safely collaborating on improved facial animation models, which would lead to improved productions. Collaboration and data re-use has been identified as an important aspect of the future of the animation industry (Liang et al. 2016), which supports our federated learning approach.

Video game studios have been criticised when it comes to surveillance and safeguarding of users (Lin and Latoschik 2022; Kröger et al. 2023). We note that while our approach allows secure training of AI models on driver data, it does not reduce the risk to drivers of data misuse by the managers themselves, who must still comply with all the relevant regulation and contractual obligations.

3.3 Driving Facial Animation Systems

Drivers of facial animation systems are defined to be the individual whose data is used to drive the animation of a trained model. The inputs to state-of-the-art facial animation systems are almost always sensitive, because they include personally-identifying image and/or audio data from drivers. Drivers includes both:

- *interacting drivers*: players of video games, who are interacting with the software and may not be aware their data could be stored and used by trainers, and
- *consenting drivers*: voice actors, who give explicit consent for their data to be used.

Drivers have the least agency of the three interest-holders, therefore much legislation is directed towards protecting the rights of drivers, such as GDPR.

The key threat to drivers is represented by managers. For example, video game studios can store sensor input and use it for non-game-related predictions about the driver like health insurance (Kröger et al. 2023). Interacting and consenting drivers have different privacy demands. Voice actors for example, give explicit consent for their data to be used for animation purposes, and in some cases, AI model training. However, in an online video game or other public lobby, interacting drivers may wish to have their identity obscured, and remain private (Eltanbouly, Halabi, and Qadir 2024).

Our system helps ensure that managers do not share driver data with third parties such as trainers, but it does not stop the managers themselves from misusing such data. Proper contracts must be drawn up with consenting drivers that allow AI model training, for example. To better protect interacting drivers, extending our silo-FL approach to device-level FL would allow driver data never to leave the gaming device or console, but this is beyond the scope of the present work.

3.4 Threat Model

Our threat model considers a “curious” central server, who wishes to uncover information about local client data, but does not break the training protocol. We do not consider malicious clients, model poisoning attacks, or misuse of the trained animation system. While we have identified managers as threats to drivers, this is a governance and contractual matter rather than one our model architecture can address, and is out of scope for this work.

4 Method

4.1 Model Architecture

We build upon Imitator (Thambiraja et al. 2023), a transformer-based audio-driven facial animation model. The model takes audio as input and aims to predict a sequence of visemes (canonical mouth shapes which correspond to speech sounds). Concretely, with reference to Figure 3, the audio input is passed into a wav2vec2 model (Baevski et al. 2020), producing audio embedding $\hat{a}_{1:T}$. This is decoded by an auto-regressive viseme decoder made up of a positional encoding, a self-attention and cross-modal attention layer, and a feedforward layer which outputs a character-agnostic motion feature $\hat{v}_{1:T}$. A style-adaptable motion decoder, made up of linear layers with ReLU activations, maps these features to person-specific facial expressions $\hat{y}_{1:T}$ in the form of vertex displacements to a template mesh.

The speaker identity vector is added to the output frames of the viseme decoder, thus making the viseme decoder identity agnostic. This means that Imitator can be efficiently finetuned for a new target identity with most of the model frozen, see Figure 4.

We train with a supervised loss adapted from the original Imitator work (Thambiraja et al. 2023), combining a vertex reconstruction term and a velocity term that promotes tem-

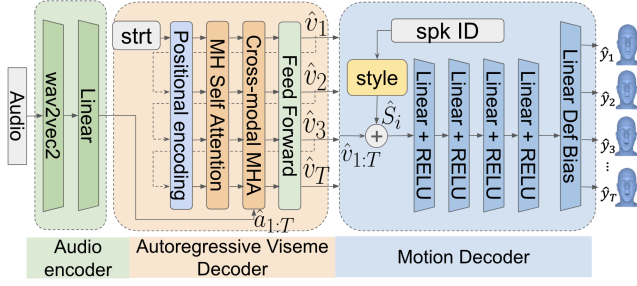


Figure 3: The Imitator model architecture (Thambiraja et al. 2023).

poral consistency:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}}, \quad (1)$$

with weights $\lambda_{\text{MSE}} = 1.0$ and $\lambda_{\text{vel}} = 10.0$. Following (Fan et al. 2022), the model is trained autoregressively rather than with teacher forcing. For V training sequences with T_v frames each, ground-truth vertices $y_{t,v}$ and predictions $\hat{y}_{t,v}$, the two terms are

$$\mathcal{L}_{\text{MSE}} = \sum_{v=1}^V \sum_{t=1}^{T_v} \|y_{t,v} - \hat{y}_{t,v}\|^2, \quad (2)$$

$$\mathcal{L}_{\text{vel}} = \sum_{v=1}^V \sum_{t=2}^{T_v} \|(y_{t,v} - y_{t-1,v}) - (\hat{y}_{t,v} - \hat{y}_{t-1,v})\|^2. \quad (3)$$

We omit the bilabial-closure term used by (Thambiraja et al. 2023), as HDTF does not have the required time-aligned transcripts.

Several architectural features make the Imitator model suitable for our privacy extensions. Firstly, as the wav2vec2 audio encoder is pre-trained, it can be frozen to avoid encoding information about sensitive training speakers. Secondly, disentangling identity-specific style information from generic viseme decoding suits the application of silo FL, as unique speaking styles can be kept private from the central server as described in Section 4.2. Finally, the model takes up a relatively small memory footprint, with wav2vec2 comprising the majority of the parameters as shown in Table 1, which means that by freezing wav2vec2 during federated training, we dramatically reduce the communication overhead; one of FL’s greatest challenges.

4.2 Privacy Architecture

Our approach makes three privacy modifications to the Imitator model to address three areas of information leakage. Freezing the audio encoder stops training speaker information from being learned, not sharing the speaker identity embedding eliminates this explicit identity encoding from being shared, and secure aggregation protects shared updates in transit. The result is that the updates that are communicated with the central server contain minimal speaker information as shown in Section 6.

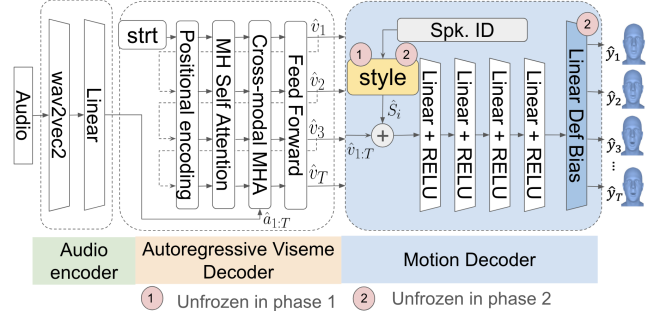


Figure 4: Finetuning step for a new identity in Imitator.

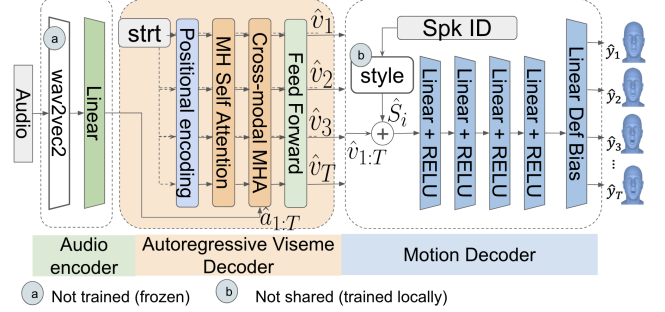


Figure 5: Private training for Imitator (default for all FL settings).

Parameter Freezing We use wav2vec2 in our audio encoder layer because it is pretrained in a self-supervised manner on public audio (LibriSpeech (Panayotov et al. 2015)) to extract useful speech representations for many downstream tasks. The Imitator authors train the whole model end-to-end, meaning the wav2vec2 layers encode speaker-identifying characteristics, as we show in Section 6.1. We therefore freeze wav2vec2 during private FL training, but keep it trainable during the public pretraining step to match the baseline setup.

During the private federated learning phase we freeze wav2vec2, resulting in a total of 1.1M shared trainable params as shown in Table 1. This reduces the per-round communication by $86\times$ versus naive FL over the full model, and no private gradients backpropagate through the audio encoder. We assess the privacy risks of training wav2vec2 in Table 6, and assess the impact of freezing on performance in the supplementary material.

Private Speaker Identity The second privacy modification is to keep speaker identity matrix weights private and not communicate them with the central server at the aggregation step.

We find that we can safely share the motion decoder with the central server, despite the fact that these layers receive as input the identity specific column of the speaker identity matrix. We show this in Section 6, specifically Table 5.

Secure Aggregation The third privacy modification is to train with secure aggregation (SA) (Bonawitz et al. 2016): masking the individual client updates with pairwise pseudo-

Component	Parameters	% of total
wav2vec2 (frozen in FL)	94,371,712	98.85%
Viseme decoder	50,240	0.05%
Speaker identity matrix	64N	(varies)
Motion decoder	996,125	1.04%
Total ($N = 8$)	95,467,805	100%
Trainable in FL	1,095,581 + 64N	$\approx 1.15\%$

Table 1: Parameter count by model component, for a model with N local speakers in the identity matrix.

random masks derived from shared seeds between each pair of clients (s_{ij}). Client i 's aggregate mask is shown in Equation 4

$$\sum_{j>i} \text{PRG}(s_{ij}) - \sum_{j<i} \text{PRG}(s_{ij}) \quad (4)$$

So the sum of all client masks is zero, and the aggregated model is mathematically equivalent to the aggregation of unmasked clients. We also show this empirically in the supplementary material.

4.3 Training Pipeline

Our training pipeline consists of three steps: 1. public pre-training 2. private finetuning via FL 3. adaptation to new identities.

Public Pre-Train We pretrain the model on VOCASET (Cudeiro et al. 2019). For more details on the dataset see Section 5. For this phase, the entire model is trainable, including the wav2vec2 audio encoder, and data is held centrally. This reproduces the baseline Imitator model (Thamiraja et al. 2023), achieving similar post-adaptation performance on the VOCASET test speakers: 4.07 mm vs 3.78 mm Lip Max and 0.107 mm vs 0.090 mm Lip L2 (although not a perfect replication as we note in Section 8). This public training step serves two purposes: firstly, it simulates a facial animation provider establishing a base model from publicly available (centralised) data, collected with explicit research consent. Second, it adapts the audio encoder to produce better representations for the task of facial animation before the private FL phase, where wav2vec2 is frozen. Note the finetuning phase is where this paper makes its privacy contribution, as this is where the data must not be distributed.

Federated Learning We train the shared model components with FedAvg (McMahan et al. 2017a). At round t , each selected client starts from the current global weights w_t , trains the model for several epochs on its own speaker data, and returns only the updated shared parameters. The server computes an average of client updates, weighted by local data size:

$$w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k \quad (5)$$

where w_{t+1}^k is client k 's model after local optimisation, n_k is the number of local examples, and $n = \sum_k n_k$ over the selected clients. The data weighting is therefore the factor $\frac{n_k}{n}$; the learning rate is applied inside each client's local

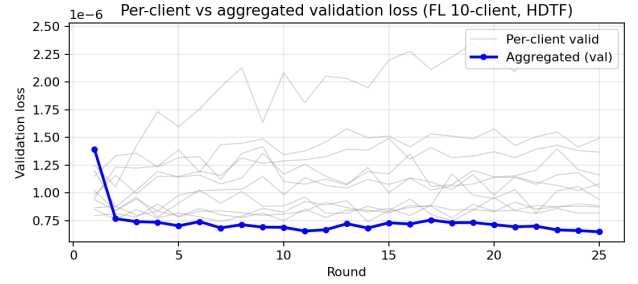


Figure 6: Per-client validation loss (grey) vs aggregated validation loss (blue) across FL rounds.

optimiser rather than as a separate server-side step. Our experiments are non-IID by construction: each client's local dataset contains all training clips for one or more speakers, with no speaker overlap across clients.

Private Fine-Tuning We adapt Imitator for training in a FL setting across multiple client devices with non-IID local datasets. First, the publicly pretrained model described in 4.3 is initialised on each device for all weights except the character-specific speaker identity matrix which is randomly initialised (Kaiming uniform). After each round (of several local epochs) we perform the aggregation step using FedAvg with secure aggregation (described in Section 4.2) to update the central model before updating each device with this new model. Two elements are not shared with the central server: The wav2vec2 audio encoder which is frozen and not updated throughout FL training; and the speaker ID matrix which is kept private. Each client's speaker ID matrix persists across rounds, so style learning accumulates locally while the motion decoder improves through collaboration. Despite the motion decoder learning to output speaker-specific styles, we find that speaker style identity information is protected from an attacker who has access to the motion decoder, but not the speaker identity matrix, as we show in Section 6.1.

We train in the federated setting for 15 rounds of 10 local epochs. Figure 6 displays the per-round valid loss, showing how individual client curves are noisy and in some cases increase each round, which would normally suggest learning is not occurring, but the aggregated model valid loss decreases relatively smoothly.

Adaptation to New Identities In line with the original paper, we freeze all layers except for the final motion decoder layer and the speaker identity matrix, as depicted in Figure 4. Specifically, we instantiate a new column in the matrix for the target speaker, initialised as the mean of the trained obj_vector columns plus small Gaussian noise (std = $0.1 \times$ base std). Stage 1 trains only this new column (100 epochs); Stage 2 additionally unfreezes the motion decoder's final output layer (100 epochs). We train on just ~ 12 – 20 seconds of data, in line with the original implementation.

5 Experiments

We train our models on two datasets, HDTF (Zhang et al. 2021) and VOCASET (Cudeiro et al. 2019). VOCASET is a high-quality dataset of studio-captured 4D facial scans, consisting of 12 speakers, each recorded on 40 sentences of 3–4 seconds each ($\sim 2\text{--}3$ min per speaker, ~ 30 min total), and aligned to the FLAME mesh (5023 vertices $\times$ 3). We apply the same split as Imitator: 8 train / 2 valid / 2 test speakers. HDTF data is extracted using MICA (Zielonka, Bolkart, and Thies 2022) in line with (Saunders and Namboodiri 2024), and comprises 66 speakers, 50 train / 6 valid / 7 test; each speaker has $\sim 11\text{--}12$ cut clips of ~ 10 s each.

When discussing results, the “baseline” is the Imitator model, trained centrally on VOCASET. This is also the starting point for the public pretrain, private FL finetune model which we refer to hereafter as the “hybrid” model.

For our facial animation quality metric, we report Lip Max in line with the literature (Richard et al. 2021; Fan et al. 2022; Thambiraja et al. 2023), which is computed for each frame as the maximal ℓ_2 error of all lip vertices, then averaged across all frames in the test set. Where stated we also report Lip L2, the *mean* (rather than maximal) ℓ_2 error over lip vertices per frame, averaged across frames. We also show qualitative examples of our best model compared with the Imitator baseline in Figure 7.

We evaluate all models after adapting them to unseen target identities. This is the most relevant use-case for our setting where a studio would need to quickly adapt the model to a new speaker or character style from a limited amount of data.

5.1 Federated Learning Results

Table 2 shows these results for the VOCASET centralised baseline model, a model trained only on HDTF (both in federated and centralised settings), and the hybrid model (public pretrain on VOCASET, private FL finetuning on HDTF). All values are Lip Max (mm) averaged across $n=3$ independent training seeds. The headline finding is that the hybrid models outperform the VOCASET baseline on unseen speakers from HDTF, with the 10-client model improving Lip Max from 4.83mm to 4.59mm, $\approx 12\times$ the error bar, without impacting performance vs the baseline on the VOCASET test set (4.08mm vs 4.07mm). The hybrid 50-client model still outperforms the baseline despite only having one speaker’s data on each silo. We also observe that the HDTF-only models underperform the baseline on both test sets, even when trained centrally, likely due to the lower quality of the HDTF data compared to VOCASET (MICA reconstructions vs studio 4D scans). This strengthens our hybrid model result: even this noisy data improves model performance in a private FL setting.

Figure 7 shows the per-vertex ℓ_2 error as a heatmap on the FLAME mesh. We show the baseline Imitator trained on VOCASET with the Hybrid FL 10-client model on one utterance from a held-out speaker from VOCASET test set and one from the HDTF test set. Blue sections show close alignment with the ground truth, while red indicates large ℓ_2 vertex error. These qualitative examples show some improvement for the FL approach in the lip area for the HDTF

Model	VOCASET (mm)	HDTF (mm)
Baseline	4.07 $\pm$ 0.04	4.83 $\pm$ 0.01
HDTF centralised	5.57 $\pm$ 0.02	5.36 $\pm$ 0.004
HDTF FL 10-client	5.37 $\pm$ 0.05	5.03 $\pm$ 0.03
HDTF FL 50-client	5.61 $\pm$ 0.05	5.36 $\pm$ 0.03
Hybrid 10-client	4.08 $\pm$ 0.02	4.59 $\pm$ 0.02
Hybrid 50-client	4.06 $\pm$ 0.02	4.69 $\pm$ 0.02

Table 2: Style adaptation to unseen test speakers: VOCASET $n=2$, HDTF $n=7$. “HDTF” models are trained from scratch on HDTF. “Hybrid” models are the baseline model privately-finetuned with FL on HDTF. Values are mean Lip Max (mm) across speakers, averaged over 3 independent training seeds, $\pm$ sample standard deviation. Lower is better.

Tag	σ	VOCASET	HDTF
uniform (baseline)	0.00	5.37	5.03
het-low	1.10	5.36	5.03
het-mid	1.67	5.36	5.02
het-high	2.45	5.34	5.00
het-dominant	3.35	5.39	5.03

Table 3: Effect of speaker-count heterogeneity across 10 FL clients on HDTF. σ is the standard deviation of per-client speaker counts. Lip Max (mm) reported.

speaker, and roughly equal performance on the VOCASET speaker, as evidenced by the quantitative results.

5.2 FL Client Heterogeneity

In Table 2, we show results with different client counts (10 and 50). To represent more realistic scenarios for FL, here we vary level of data heterogeneity σ . We show that even moderate heterogeneity doesn’t appear to result in any performance degradation, which suggests our results are robust to more real-world situations where small games studios do not have equal-sized private datasets.

5.3 Secure Aggregation

We show that SA training is provably lossless in Section 4.2, so for increased training speed, most of our FL models are trained without secure aggregation. However, for further empirical proof, we finetune the full 50-client model on HDTF with secure aggregation, finding no difference to the trained model (results in supplementary material).

We note that SA prevents the server from observing individual client contributions, but the resulting aggregated model can leak information about those contributions. To formally bound this leakage, we investigate user-level DP in Section 6.4, finding it is not deployable at this scale. We instead empirically evaluate privacy risks to the aggregated model through speaker-ID and membership inference attacks (Section 6) and our gradient inversion attack conservatively assumes access to a raw, unmasked client update, which should not theoretically be possible due to SA.

SA imposes a significant computational cost from mask generation. Table 4 reports per-round wall-clock timings.

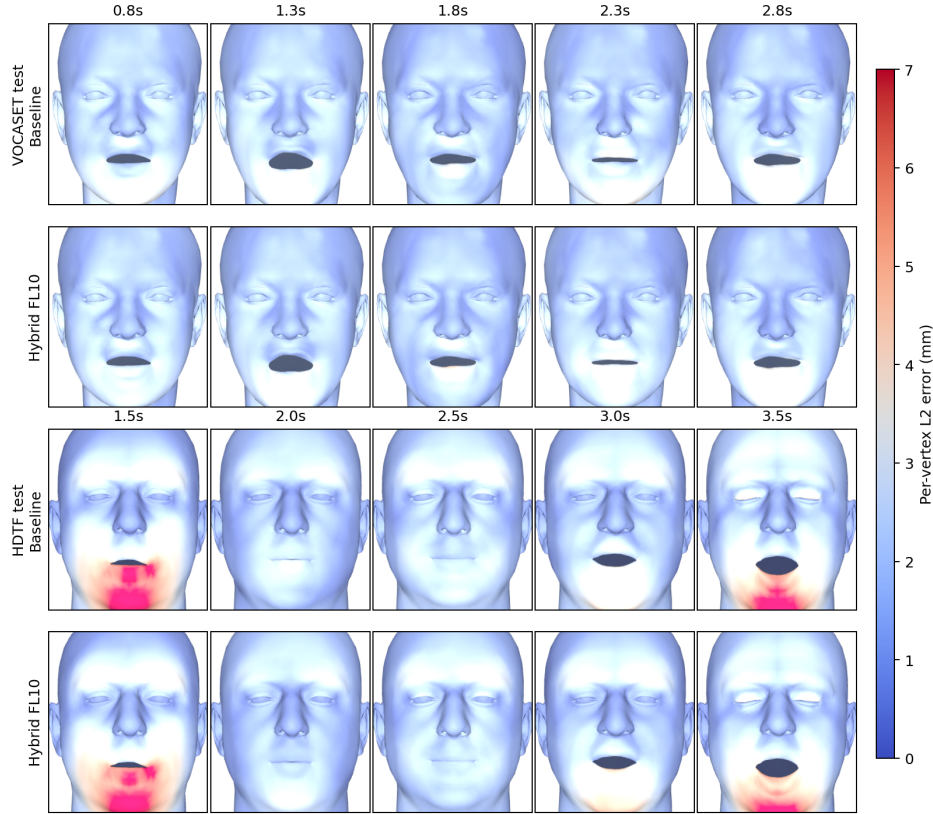


Figure 7: Per-vertex ℓ_2 prediction error (mm) for baseline and hybrid models compared to ground truth. Top: a held-out VOCASET test speaker; bottom: a held-out HDTF test speaker. Frames are selected at 0.5 s intervals. Our model preserves or improves quality while improving privacy. Zoom in for better comparison.

Phase	FedAvg	SA
Client training	$\sim 960s$	$\sim 960s$
Aggregation (sequential sim)	0.1s	2530s
Validation	9s	9s
Total per round	$\sim 970s$	$\sim 3500s$
Total (15 rounds)	$\sim 4h$	$\sim 14.6h$

Table 4: Per-round timings for FedAvg vs. SA at 50 clients.

In our single-machine simulation, mask generation is performed sequentially for all 50 clients, each generating 49 pairwise masks over $\sim 1M$ shared parameters. In realistic deployment, the slowdown would not be so severe as clients perform this work in parallel. The per-client overhead is approximately the reported figure divided by the client count.

6 Privacy Evaluation

In this section we evaluate the privacy gains of our model architecture using three feasible attack vectors, each with different motivations and implications for the interest-holders introduced in Section 3. We also include a study on the feasibility of differential privacy in this setting.

6.1 Speaker-ID Attack

To show how much speaker information leaks from model-internal layers, we train a 2-layer MLP with 256 and 128 hidden dimensions respectively and ReLU activation, before a final output layer with output dimensions = n_{speakers} . We pass samples from the train and test data through the model, and extract the output of four model-internal layers for classification:

- L0: Audio encoder output.
- L1: Viseme Decoder output: transformer’s audio-driven mouth-shape features, before style is added. Speaker-agnostic by design.
- L2: Motion Decoder input: Viseme decoder output with the per-speaker style added.
- L3: Final motion decoder output (before adding to mesh)

We train the speaker identification model for 200 epochs with an Adam optimiser and a learning rate of $1e-3$. The training data is a 70/30 speaker-level train/test split of the extracted hidden representations. We extract hidden states from all HDTF training-speaker samples (~ 800 samples for 50 speakers), and train one speaker ID model for each of the four layers.

Table 5 reports the results: zeroing the obj_vector collapses classifier accuracy at style-conditioned layers (L2, L3

Model	speaker vector	L0	L1	L2	L3
FL 10-client	real	0.07	0.05	1.00	0.48
FL 10-client	zero	0.10	0.04	0.02	0.02

Table 5: Speaker-ID classifier accuracy on internal hidden states of FL-trained Imitator models (50 HDTF train speakers, chance=2%).

Model	MIA AUC	EER
Centralised (unfrozen wav2vec2)	0.9314	0.1429
Centralised (frozen wav2vec2)	0.6143	0.4000

Table 6: Speaker-level MIA AUC-ROC and EER on centralised HDTF models, 0.5 = chance.

pre-template) to chance, confirming that without access to the private per-speaker column of the identity matrix, the shared decoder reveals nothing about speaker identity. This property holds even for 50 clients, 1 speaker per client. See these results in the supplementary material.

6.2 Membership Inference Attack

To test whether an attacker can identify which speakers were used for training from just the model outputs, we apply a loss-based speaker-level MIA (Yeom et al. 2018).

For each HDTF training speaker and held-out test speaker, we compute the model loss over that speaker’s clips and use average loss as the attack score. Because only training speakers have learned style IDs, we use a worst-case “best-match” attacker: for every sample, the attacker evaluates all available training-speaker style IDs and keeps the lowest loss. We report AUC-ROC and equal error rate (EER), where AUC=0.5 corresponds to random guessing. The intuition is that the model should have lower loss on learned training samples, making membership inference possible.

The results are presented in Table 6. Without freezing the wav2vec2 audio encoder, the attacker reaches near-perfect AUC=0.93, meaning almost every training speaker is ranked as more likely a member of the training set than any held-out speaker. Freezing wav2vec2 during training reduces AUC to 0.61, under the same best-match MIA, much closer to random guessing.

We measure MIA on centralised models as an upper bound on the leakage an unfrozen audio encoder would expose, which motivates our decision to freeze wav2vec2 during FL training.

This result does not constitute a formal privacy guarantee, but it shows that training the audio encoder on private speech considerably increases membership leakage. The formal route to protecting membership is user-level DP, which we evaluate in Section 6.4.

6.3 Gradient Inversion Attack

We investigate a gradient inversion attack (Geiping et al. 2020), where an attacker can reconstruct training data samples from the updates communicated with the central server. Our secure aggregation protocol should make it impossible for an attacker to acquire the model updates in the first place,

but we still evaluate the gradient inversion attack on the unmasked weight updates for completeness.

Formally, given a gradient $g = \nabla_{\theta} L(\theta; X, Y)$ leaked from a client, the attacker solves

$$X^* = \arg \min_{X'} 1 - \frac{\langle \nabla_{\theta} L(\theta; X', Y), g \rangle}{\|\nabla_{\theta} L(\theta; X', Y)\| \|g\|} + \alpha \cdot \text{TV}(X') \quad (6)$$

where the first term is the cosine distance between the observed gradient and the gradient produced by a dummy input X' , and $\text{TV}(\cdot)$ is a total-variation regulariser (weight α) that encourages temporal smoothness in the recovered vertex sequence. The labels Y (per-frame template mesh and conditioning audio) are assumed known to the attacker. We optimise X^* for T Adam steps.

Using a single unmasked FL client update, we optimise dummy vertex sequences to match its gradients. We find that the attack completely fails: reconstruction ℓ_2 error is $\approx 1596\text{mm}$ vs 108mm random baseline ($\sim 0.07 \times$ random noise). It is possible that the high-dimensionality of the vertex output space makes this attack very difficult to optimise.

6.4 Differential Privacy Study

To investigate whether a formal privacy guarantee is achievable at our scale, we implement user-level DP. We apply per-client update clipping plus Gaussian noise at aggregation.

At round t , each client k computes its update $\Delta_t^k = w_t^k - w_{t-1}$ and clips it to a fixed ℓ_2 ball,

$$\tilde{\Delta}_t^k = \Delta_t^k \cdot \min\left(1, \frac{C}{\|\Delta_t^k\|_2}\right). \quad (7)$$

The server then aggregates with calibrated Gaussian noise,

$$w_t = w_{t-1} + \frac{1}{K} \sum_{k=1}^K \tilde{\Delta}_t^k + \mathcal{N}\left(0, \frac{\sigma^2 C^2}{K^2} I\right), \quad (8)$$

where C is the clipping norm, K the number of clients, and σ the noise multiplier. The corresponding (ϵ, δ) guarantee at $\delta = 10^{-5}$ over $T = 10$ FL rounds is computed via Rényi DP accounting (Mironov 2017).

Table 7 shows that DP destroys model quality on the 1.1M-parameter vertex model: the best configuration ($\sigma = 0.5$, $K = 10$, $C = 50$) produces Lip Max error in the billions of millimetres. To simplify the task, we try predicting 103-dimensional FLAME parameters instead of 15069-dimensional vertices, reducing shared trainable parameters by $\sim 9 \times$ (Table 8). Even this smaller model fails to learn under DP: at $\sigma = 0.5$ with 50 clients ($\epsilon \approx 5.7$), loss is worse than random initialisation.

These results motivate our private FL architecture of frozen wav2vec2, private style embeddings and secure aggregation over formal DP. We show that at realistic cross-silo scale (10–50 clients), DP is not deployable, whereas our approach provides strong empirical privacy at zero quality cost.

7 Discussion

As training datasets scale, the issue of privacy and data governance grows larger. For managers of facial animation

Clients	σ	ε	noise/signal	Viable?
10	0.5	12.1	$33\times$	No
50	0.5	5.7	$\sim 7\times$	No
<i>100</i>	<i>0.5</i>	<i>4.4</i>	$\sim 3\times$	Borderline
500	0.5	2.8	$\sim 0.7\times$	Yes

Table 7: User-level DP analysis for the vertex model (1.1M shared parameters). Values in *italics* are extrapolated and were not tested as our maximum available clients was 50. ε computed via RDP accounting at 10 FL rounds.

Config	σ	ε	Loss
Baseline (no DP)	-	-	0.061
Random init	-	-	0.067
DP 50-client	0.5	5.7	0.079–0.099

Table 8: DP on the FLAME parameter model (103-dim output, 123K shared parameters, 50 clients).

systems, choosing what data is and isn’t shared with a central model provider matters. This represents control over sensitive biometric data, stored in accordance with GDPR and contractual commitments, and business-critical assets such as character designs and information about upcoming releases. Our speaker-ID result shows that without the speaker ID matrix, an attacker cannot predict character-specific speaking styles from the model weights, protecting these critical assets.

For trainers, more data produces better models, but data collection and storage presents technical and regulatory challenges. Our system allows trainers to improve their model, while managers retain data sovereignty. Our approach hides the parts of the model which may learn speaker-identifying representations from the central server. The clients therefore only provide useful gradient information for the identity-agnostic model layers (the viseme decoder and motion decoder), which is sufficient to improve generalisation to unseen identities, ultimately improving model performance which is the trainers’ priority. Our system architecture moves raw data off the trainer’s premises, but we acknowledge that power imbalances still remain, as the trainer retains the model, the pretrained checkpoint, the protocol, and the question of who gets to be a client.

Drivers of facial animation systems are the least supported by our implementation. We have made a positive privacy contribution for drivers, freezing wav2vec2 reduces membership inference AUC from 0.93 to 0.61, the speaker identity matrix prevents style identification from shared layers, and gradient inversion attacks fail. But these are leakage reductions, not consent. In our cross-silo FL setting, the data is owned and controlled by managers. For consenting drivers, their contracts must explicitly allow use of performances for model training. To protect interacting drivers, however, more work would be required such as extending to device-level FL with formal guarantees from differential privacy.

8 Limitations

We acknowledge several limitations to this work. On experimental design, we only simulate the federated learning setting on a single device. We experiment with different numbers of clients and data splits, but do not consider client dropout, malicious clients, or adversarial server settings.

Our HDTF mesh data is fitted from monocular video, and is not studio-quality like VOCASET. This is a limitation of the dataset but a useful test of model robustness to noisy data.

We acknowledge with the exception of secure aggregation, our privacy contributions are empirical and model-specific. We have shown that differential privacy is not applicable to collaboration at this scale, and have proposed what scale of data may be required for DP to become feasible for our specific setting.

The speaker-ID analysis is limited to a learned MLP probe. We show our architecture is strong against this type of attack, but we cannot rule out an attacker leveraging other speaker or style-related information encoded in temporal dynamics or other information we do not consider.

Our approach reduces data centralisation and allows studios to retain sovereignty over their local datasets, but does not protect against misuse of actor and player data by the studios themselves.

We aim to reproduce the Imitator baseline as closely as possible, but our baseline result falls slightly short of the original work: 4.07 mm vs 3.78 mm Lip Max. This is likely due to our different speaker identity matrix initialisation (we take the average instead of exhaustive best-match search), and we drop the bilabial closure loss term as we lack the required transcripts for HDTF.

We also acknowledge there are no user studies in this paper, which can be more useful to assess perceptual model performance than quantitative metrics for facial animation.

9 Future Work

Because we improve performance and data privacy of our facial animation model for the FLAME 3DMM, a natural extension to our method would be to apply some texture to the FLAME mesh for fine-grained detail, through an approach like gaussian splatting or diffusion.

The approach transfers to other tasks where pretrained features and disentangled identity representations exist, e.g. speech synthesis. We release our code to evaluate such uses.

For extensions to our federated learning approach, deploying the system in a real multi-studio setting with stragglers, drop-out, and malicious clients would be a worthy contribution. Our cross-silo FL approach could be extended to cross-device FL over individual player voice data, aligned with monocular video feed for a training signal. Such a case would require robust privacy guarantees; at that scale, user-level DP may become viable, and future work could verify this by extending our DP study with an order of magnitude more clients to achieve a formal privacy guarantee.

Acknowledgements

Jack Parry is supported by a University Research Studentship Award (URSA) at the University of Bath.

Position Statement

This work was motivated by the lead author's prior experience as an ML engineer at a facial animation provider serving the games industry. That direct exposure informs the institutional setting we describe in Section 3: the actors, incentives, contractual constraints, and what a deployable system looks like in practice. This also presents a potentially biased position, given the facial animation provider in this setting stands to gain from broader training data without the legal requirements of data custody. This research is not industry-funded, however, and was conducted in a university research setting. This work is grounded in a machine learning research perspective, representing a technical approach to data privacy and governance, rather than as a legal or regulatory framework, or from a workers rights advocacy perspective. We did not consult voice actors or studio QA artists whose work stands to be affected by such a system.

Ethical Statement

All experimental data used in this paper comes from public datasets (VOCASET and HDTF), and no new biometric data was collected. We present hypothetical industry scenarios, such as where QA artists may hand-edit model outputs as training data, but no hand-animation was carried out; all our training data is either publicly available (VOCASET) or generated with automatic techniques (HDTF, via MICA). Our paper does not include a user study of perceived animation quality, which we acknowledge as a methodological limitation in Section 8.

Broader Impact

Audio-driven facial animation technology can be misused. Improvements in lip-sync quality and identity generalisation, including those obtained via privacy-preserving collaboration, also present opportunities for identity impersonation ("deepfake" creation). Our paper does not address inference-time misuse, we do not claim that our privacy architecture is a solution to this.

Risk of Overstating Privacy Guarantees

As we make clear in the paper, our privacy contributions are empirical and not formal guarantees. MIA AUC drops from 0.93 to 0.61 under frozen wav2vec2, but this is still not random guessing. Any studio or provider deploying this framework should not cite it as evidence that performer privacy is being fully preserved, rather just measurably increased.

Performer Consent Is Bypassed

The framework treats the studio as the data holder. Voice actors and motion-capture performers can have their performances used to train shared facial animation models without consent specifically for that purpose. We encourage adopters to negotiate AI-training consent separately rather than assuming production contracts cover it.

Barriers to Participation

Federated learning has non-trivial ML and compute requirements. Small studios may be unable to participate, which could lead to an oligopoly of studios holding outsized power, which may not represent a large improvement over the current centralised paradigm.

International Data Flows

Federated updates derived from performers' biometric data may be regulated differently in different jurisdictions, and our paper does not provide guidance on how to ensure compliance with differing global regulations.

References

- AbdulRahman, S.; Tout, H.; Ould-Slimane, H.; Mourad, A.; Talhi, C.; and Guizani, M. 2020. A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. *IEEE Internet of Things Journal*, 8(7): 5476–5497.
- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Aggarwal, D.; Zhou, J.; and Jain, A. K. 2021. Fedface: Collaborative learning of face recognition model. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 1–8. IEEE.
- Andrew, G.; Thakkar, O.; McMahan, B.; and Ramaswamy, S. 2021. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34: 17455–17466.
- Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.
- Bell, J. H.; Bonawitz, K. A.; Gascón, A.; Lepoint, T.; and Raykova, M. 2020. Secure single-server aggregation with (poly) logarithmic overhead. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 1253–1269.
- Berchtold, W.; Schäfer, M.; Liu, H.; Takahashi, F. T.; Schmitz, A.; Zmudzinski, S.; Steinebach, M.; and Wieneke, J. 2013. Video game watermarking. In *Media Watermarking, Security, and Forensics 2013*, volume 8665, 33–46. SPIE.
- Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2016. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Chen, C.; Xu, H.; Wang, W.; Li, B.; Li, B.; Chen, L.; and Zhang, G. 2021. Communication-Efficient Federated Learning with Adaptive Parameter Freezing. In *2021 IEEE 41st*

International Conference on Distributed Computing Systems (ICDCS), 1–11.

Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 10101–10111.

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.

Ding, Z.; Zhang, X.; Xia, Z.; Jebe, L.; Tu, Z.; and Zhang, X. 2023. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12736–12746.

Edwards, P.; Landreth, C.; Popławski, M.; Malinowski, R.; Watling, S.; Fiume, E.; and Singh, K. 2020. Jali-driven expressive facial animation and multilingual speech in cyberpunk 2077. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks*, 1–2.

Eltanbouly, S.; Halabi, O.; and Qadir, J. 2024. Avatar privacy challenges in the metaverse: A comprehensive review and future directions. *International Journal of Human-Computer Interaction*, 1–18.

EU AI Act. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L series. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33: 16937–16947.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5784–5794.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Huang, C.; Huang, J.; and Liu, X. 2022. Cross-silo federated learning: Challenges and opportunities. *arXiv preprint arXiv:2206.12949*.

Illinois General Assembly. 2008. Biometric Information Privacy Act. 740 ILCS 14/1 et seq. Public Act 95-994, effective October 3, 2008.

Kröger, J. L.; Raschke, P.; Campbell, J. P.; and Ullrich, S. 2023. Surveilling the gamers: Privacy impacts of the video game industry. *Entertainment Computing*, 44: 100537.

Li, J.; Khiredkar, R.; He, C.; Jiang, Z.; Nam, G.; Yang, L.; Lee, J.; Zakharov, E.; Su, Z.; Abdrashitov, R.; et al. 2026. Large-scale Codec Avatars: The Unreasonable Effectiveness of Large-scale Avatar Pretraining. *arXiv preprint arXiv:2604.02320*.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6): 194:1–194:17.

Liang, H.; Sit, J.; Chang, J.; and Zhang, J. J. 2016. Computer animation data management: Review of evolution phases and emerging issues. *International journal of information management*, 36(6): 1089–1100.

Lin, J.; and Latoschik, M. E. 2022. Digital body, identity and privacy in social virtual reality: A systematic review. *Frontiers in Virtual Reality*, 3: 974652.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017a. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

McMahan, H. B.; Ramage, D.; Talwar, K.; and Zhang, L. 2017b. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.

Meng, Q.; Zhou, F.; Ren, H.; Feng, T.; Liu, G.; and Lin, Y. 2022. Improving federated learning face recognition via privacy-agnostic clusters. *arXiv preprint arXiv:2201.12467*.

Mironov, I. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, 263–275. IEEE.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Protection, D. 2018. General data protection regulation. *Intersoft Consulting*, Accessed in October, 24(1).

Qiu, L.; Gu, X.; Li, P.; Zuo, Q.; Shen, W.; Zhang, J.; Qiu, K.; Yuan, W.; Chen, G.; Dong, Z.; et al. 2025. LHM: Large Animatable Human Reconstruction Model for Single Image to 3D in Seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14184–14194.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Richard, A.; Zollhöfer, M.; Wen, Y.; De la Torre, F.; and Sheikh, Y. 2021. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1173–1182.

Saunders, J.; and Nambodiri, V. 2024. Dubbing for everyone: Data-efficient visual dubbing using neural rendering priors. *arXiv preprint arXiv:2401.06126*.

Shome, D.; and Kar, T. 2021. FedAffect: Few-shot federated learning for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4168–4175.

- 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.
- Thambiraja, B.; Habibie, I.; Aliakbarian, S.; Cosker, D.; Theobalt, C.; and Thies, J. 2023. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20621–20631.
- Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64: 131–148.
- Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023. 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. 2024a. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37: 660–684.
- Xu, Y.; Chen, B.; Li, Z.; Zhang, H.; Wang, L.; Zheng, Z.; and Liu, Y. 2024b. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.
- Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, 268–282. IEEE.
- 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*, 3661–3670.
- Zielonka, W.; Bolkart, T.; and Thies, J. 2022. Towards metrical reconstruction of human faces. In *European conference on computer vision*, 250–269. Springer.
- Zielonka, W.; Bolkart, T.; and Thies, J. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4574–4584.