

Retaining Emotions, Removing Identity: Valence–Arousal Guided De-Identification for Educational Applications

Ashwin T S
ashwindixit9@gmail.com
Vanderbilt University
Nashville, Tennessee, USA

Nihar Sanda
nihar.sanda@gmail.com
Vanderbilt University
Nashville, Tennessee, USA

Austin Coursey
austin.c.coursey@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Vaibhav Gupta
vaibhav.gupta@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Gautam Biswas
gautam.biswas@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Abstract

The integration of artificial intelligence (AI) for multimodal data analysis has enabled deep learning models to analyze students' facial expressions for emotion recognition in classroom settings. However, collecting facial data from minors raises substantial privacy and ethical concerns, necessitating privacy-preserving de-identification techniques that retain affective information. Existing methods primarily target identity suppression but often fail to preserve fine-grained emotional expressions, limiting applicability in affective computing. We present a privacy-preserving face de-identification method that replaces real identities with synthetic faces selected via valence–arousal guided identity replacement in the StyleGAN latent space. This approach ensures de-identified faces retain the original emotional expressions while suppressing biometric identity. We evaluate our algorithm on a real-world classroom dataset of 40 middle school students and the publicly available DAiSEE dataset, achieving privacy preservation accuracy of 89.97% (Classroom) and 89.94% (DAiSEE), significantly outperforming conventional anonymization techniques. Additionally, the method maintains high expression fidelity, with 92.4% manual verification agreement and low Facial Action Unit (FAU) loss. These findings establish an efficient, privacy-compliant de-identification framework for affect-sensitive AI supporting classroom analytics and educational research.

CCS Concepts

• **Applied computing** → **Education**; • **Computing methodologies** → **Biometrics**; **Machine learning**; • **Security and privacy** → **Privacy protections**.

Keywords

Emotion Recognition, Privacy-Preserving AI, Face De-Identification, Affective Computing, Learning Analytics, Student Data Privacy

ACM Reference Format:

Ashwin T S, Nihar Sanda, Austin Coursey, Vaibhav Gupta, and Gautam Biswas. 2026. Retaining Emotions, Removing Identity: Valence–Arousal Guided De-Identification for Educational Applications. In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26)*, March 23–27, 2026, Thessaloniki, Greece. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3748522.3780003>

1 Introduction

The widespread adoption of artificial intelligence (AI) and deep learning has advanced facial analysis in surveillance, healthcare, and education [11, 35]. In learning environments, affective computing enables personalized instruction, engagement tracking, and adaptive feedback by analyzing students' facial expressions [22, 4, 5, 27]. However, collecting facial data from minors raises critical privacy and ethical concerns [7]. Facial images contain highly identifiable biometric features, and misuse can lead to privacy violations. With stringent regulations such as the General Data Protection Regulation (GDPR) and the Family Educational Rights and Privacy Act (FERPA), there is an urgent need for de-identification techniques that preserve affective utility [25, 6].

Emotion recognition is central to learning analytics and understanding students' self-regulated learning (SRL) processes, as students' emotional states influence motivation, engagement, and outcomes [34, 33, 12]. Accurate affect detection helps educators and AI tutoring systems identify learning struggles, detect cognitive overload, and adjust instructional strategies. Engagement monitoring relies on subtle cues, such as confusion, frustration, and boredom, that are more informative than basic emotions [9]. Consequently, any de-identification method used in education must preserve fine-grained affective expressions while protecting privacy.

Affective expressions exist on a continuous spectrum, where subtle changes in facial movements, gaze, and micro-expressions signal a variety of cognitive and emotional states. The valence–arousal model provides a continuous representation that captures these gradations [26]. These values can be mapped to categorical academic emotions such as confusion, delight, and frustration, making valence–arousal modeling particularly valuable for educational affect detection [9, 26, 2, 1]. Thus, privacy-preserving approaches must ensure de-identified faces retain original valence–arousal characteristics to maintain model accuracy.



This work is licensed under a Creative Commons Attribution 4.0 International License. SAC '26, Thessaloniki, Greece

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2294-3/2026/03

<https://doi.org/10.1145/3748522.3780003>

Traditional de-identification (e.g., blurring, pixelation, adversarial noise) removes identity but obscures expression-related details. While deep learning approaches, such as Generative Adversarial Networks (GANs) and diffusion models, have improved anonymization, they emphasize identity suppression over emotion preservation. They are often developed for security, healthcare, and surveillance [8, 19]. Existing datasets for face de-identification are largely adult-focused and overlook academic emotions central to educational research. Confusion, frustration, boredom, and delight—directly linked to learning outcomes—are underrepresented in public datasets [17]. Current models also struggle with in-the-wild classroom data, where lighting, pose, and occlusions complicate anonymization and affect detection [12].

To address these challenges, we propose a privacy-preserving face de-identification framework tailored for educational environments. Rather than modifying facial features, our approach matches the input face to a synthetic counterpart from a pre-generated StyleGAN dataset [13], ensuring the de-identified face retains the same valence–arousal expression while removing identifiable biometric information. The key contributions are:

- *Expression-Preserving Face De-Identification*: A valence–arousal guided framework that retains emotional expressions while ensuring privacy protection.
- *Seed-Based Latent Space Matching*: Selection of a synthetic counterpart exhibiting the same affective state while being visually distinct.
- *Evaluation on Real-World Educational Data*: Validation on the DAiSEE dataset [14] as well as a dataset of 40 students working in the Ecojourneys learning environment [23].

By overcoming the privacy limitations of existing techniques while maintaining high-fidelity affect detection, this work advances privacy-aware AI in education and contributes to the broader field of privacy-preserving facial analysis.

2 Background

Face de-identification seeks to protect privacy by altering identifiable facial features while preserving image utility [8]. Methods range from simple pixelation to advanced deep learning models, yet most do not explicitly preserve fine-grained emotional expressions, limiting suitability for affective computing [8, 19].

Existing de-identification techniques can be categorized into three levels [8]: (1) *Pixel-Level Approaches*: Direct pixel modifications (e.g., blurring, pixelation) effectively remove identity but obscure subtle expressions, reducing utility for affect recognition; (2) *Representation-Level Approaches*: Deep generative models (e.g., GANs, VAEs) produce synthetic faces with altered identities. Without explicit expression control, they often distort affective states; and (3) *Semantic-Level Approaches*: Methods that adjust high-level attributes (e.g., gender, age) typically overlook expression preservation and compromise affective fidelity.

Recent advances in StyleGAN-based generation demonstrate the creation of highly realistic synthetic faces while preserving some non-biometric attributes [19, 13]. Systems such as DeepPrivacy and CIAGAN anonymize faces by replacing them with GAN-generated identities, but lack mechanisms for emotion preservation [16]. Their

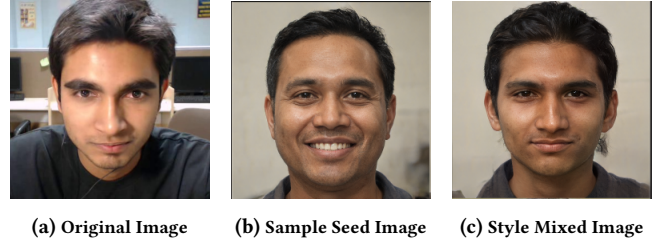


Figure 1: Example of our de-identification pipeline: (a) The original image, (b) The AI-generated seed image selected based on valence–arousal similarity, and (c) The final style-mixed de-identified image, where the facial expression from the original image is retained while identity-related features are replaced.

focus on identity suppression over expression retention limits applicability in educational affect detection. Some studies explore emotion-guided de-identification; however, they generally rely on discrete categories (e.g., happy, sad, angry) rather than continuous valence–arousal representations [21]. This discretization constrains the capture of fine-grained affective nuances critical for education-focused applications.

Despite progress, many de-identification methods alter emotional expressions during anonymization, causing loss of affective information. Addressing this gap is essential for affect-sensitive AI, particularly in educational research, where de-identification must protect privacy without distorting the emotional signals required for reliable affect detection.

3 Method

The de-identification process replaces an individual’s face with a seed image—an AI-generated face that does not correspond to any real person—thereby removing biometric identity while preserving the original affective expression. We implement this using StyleGAN [13], which provides precise control over facial attributes in a disentangled latent space [31], and TensOrgan¹. Unlike conventional methods that randomly select replacements, we introduce a valence–arousal guided seed selection mechanism to ensure the chosen seed closely matches the input’s emotional state. As illustrated in Figure 1, the original image (Figure 1a) is matched to a seed image (Figure 1b) from a precomputed dataset of AI-generated faces. The final de-identified style-mixed image (Figure 1c) is synthesized by blending facial structure from the seed while preserving the input’s affective expression.

A seed image, following prior GAN-based synthesis research, is a precomputed AI-generated face used for anonymization. Seeds are stored in a curated dataset, with each seed’s valence–arousal values extracted in advance for efficient retrieval. While valence–arousal similarity helps preserve the general emotional state, it does not guarantee structural divergence or privacy protection. To refine selection, we incorporate perceptual similarity constraints, ensuring the de-identified face is expressively accurate yet visually distinct from the original identity.

¹<https://github.com/renhaa/tensorgan>

Face Detection and Preprocessing: Each input video frame undergoes face detection and cropping using Multi-task Cascaded Convolutional Networks (MTCNN) [36]. The detected face is resized to 1024×1024 pixels to ensure compatibility with StyleGAN-based latent space projection. The cropped face is then processed using EMONet [20], a deep learning model for affect recognition, to extract its valence (positivity/negativity) and arousal (intensity) values:

$$(V_{\text{input}}, A_{\text{input}}) = \text{EMONet}(I_{\text{cropped}}) \quad (1)$$

Similarly, valence-arousal values for all seed images are precomputed, allowing for efficient search during seed selection:

$$(V_{\text{seed}}, A_{\text{seed}}) = \text{EMONet}(I_{\text{seed}}) \quad (2)$$

Valence-Arousal Guided Seed Selection: To ensure that the de-identified face retains the same emotional state, we select a seed image that minimizes the valence-arousal distance from the input image:

$$d_{VA} = \sqrt{(V_{\text{input}} - V_{\text{seed}})^2 + (A_{\text{input}} - A_{\text{seed}})^2} \quad (3)$$

To refine seed selection beyond valence-arousal similarity, we incorporate perceptual similarity constraints:

- LPIPS (Learned Perceptual Image Patch Similarity) [37]: Ensures that the structure and overall appearance remain natural.
- ArcFace Cosine Similarity [10]: Ensures the seed image is not too visually similar to the original identity to enhance privacy.

The final seed selection function is given by:

$$S_{\text{total}} = \lambda_1 d_{VA} + \lambda_2 \text{LPIPS}(I_{\text{input}}, I_{\text{seed}}) + \lambda_3 \text{ArcFace}(I_{\text{input}}, I_{\text{seed}}), \quad (4)$$

where $\lambda_1, \lambda_2, \lambda_3$ control the balance between expression retention, perceptual similarity, and identity suppression. The seed image with the lowest S_{total} is selected as the final seed.

Demographic Filtering for Age Consistency: Since all input images feature children, it is essential that the selected seed image from the seed database corresponds to the same age group. This helps prevent inconsistencies that could lead to inaccurate emotion preservation. We use a pre-trained ResNet-based age classifier [15] to filter the faces generated by StyleGAN, ensuring that only child-like seed images are chosen. The filtering process employs a probability-based threshold:

$$C(I_{\text{seed}}) = \text{child} \quad \text{if} \quad P_{\text{child}} > \tau \quad (5)$$

where P_{child} represents the probability that a face belongs to a child.

Latent Space Projection and Style Mixing Using StyleGAN: Once the most suitable seed image is identified, both the input image and the seed image are projected into the StyleGAN latent space using GAN inversion. Since real faces do not naturally exist in StyleGAN's latent space, we employ a GAN inversion technique, specifically the pSp (Pixel2Style2Pixel) encoder, to map the input face into the generator's domain before applying style mixing.

$$z_{\text{input}} = G^{-1}(I_{\text{cropped}}) \quad (6)$$

$$z_{\text{selected}} = G^{-1}(I_{\text{selected}}) \quad (7)$$

To generate a de-identified yet expression-preserving face, we use style mixing:

$$z_{\text{mixed}} = \text{mix}(z_{\text{input}}, z_{\text{selected}}), \quad (8)$$

where mix refers to StyleGAN Style Mixing, where the latent codes of two images are combined at different layers of the generator [18].

$$I_{\text{de-id}} = G(z_{\text{mixed}}) \quad (9)$$

This ensures that the biometric identity of the original face is removed, while the affective expressions remain similar to the input.

Microexpression Retention with FAU Loss: While matching valence and arousal helps maintain macro-level emotional expressions, it is crucial to retain fine-grained microexpressions to ensure that subtle movements of facial muscles remain unchanged after de-identification. We achieve this using the Affectiva SDK through iMotions, a leading affective computing platform that offers real-time detection of Facial Action Units (FAUs) based on the Facial Action Coding System (FACS).

Step 1: Extract FAU Vectors Using Affectiva SDK: The Affectiva SDK provides precomputed Facial Action Units (FAUs) for each detected face. We extract FAU embeddings for both the original and de-identified images.

$$f_{\text{input}} = \text{FAU}(I_{\text{input}}), \quad f_{\text{de-id}} = \text{FAU}(I_{\text{de-id}}) \quad (10)$$

where:

- f_{input} is the FAU vector for the input image.
- $f_{\text{de-id}}$ is the FAU vector for the de-identified image.

Step 2: Compute FAU Loss for Microexpression Retention: To ensure that microexpressions remain consistent, we minimize the L2 loss between the FAU vectors of the original face and the de-identified face:

$$\mathcal{L}_{\text{FAU}} = \sum_{i=1}^N \|f_{\text{input},i} - f_{\text{de-id},i}\|_2^2, \quad (11)$$

where:

- $f_{\text{input},i}$ is the i -th Facial Action Unit (FAU) intensity for the original face.
- $f_{\text{de-id},i}$ is the corresponding i -th FAU intensity for the de-identified face.
- N is the total number of detected FAUs.

This maintains consistent muscle movements, microexpressions, and facial details, avoiding any emotional distortion from de-identification.

Implementation Considerations and Scalability: To optimize computational efficiency, valence-arousal values for seed images are precomputed and stored in an embedding space. During inference, fast nearest-neighbor search [24] is performed for efficient seed selection.

GAN inversion and style mixing introduce computational overhead, but since face crops are processed individually rather than

full video frames, the pipeline remains scalable for real-world educational datasets.

4 Results and Discussion

4.1 Dataset Details

To evaluate the effectiveness of our privacy-preserving face de-identification methodology, we utilize two datasets: one collected in a real-world classroom learning environment and the publicly available DAiSEE dataset [14]. The classroom dataset consists of middle school students engaged in a game-based learning environment, while the DAiSEE dataset includes adult learners participating in e-learning activities. These datasets allow us to assess our method's performance across different age groups and learning contexts while adhering to privacy regulations.

Since our method involves valence-arousal guided seed selection, we did not use all available images from these datasets. The initial filtering was performed based on valence-arousal similarity, where only input images whose valence-arousal values closely matched one of the 1,024 precomputed seed images were retained. Approximately 53% of the images across both datasets met this criterion and were used for the final analysis. The remaining images, where valence-arousal differences exceeded the predefined threshold, were excluded to prevent expression mismatches. After this initial selection, we applied perceptual similarity constraints (LPIPS) and demographic filtering (age classification) to ensure that the final evaluation set preserved emotional expressions while maintaining privacy and demographic consistency. This seed-based selection introduces certain coverage limitations in the emotional spectrum, which are discussed in detail in Section 4.2.

Ecojourneys Classroom Learning Environment Dataset: The first dataset was collected as part of a game-based learning environment designed for life science education [23]. The study was approved by the Institutional Review Board (IRB) of Vanderbilt University, and informed consent was obtained from all participants. A total of 40 middle school students (22 female, 18 male) in the United States, aged 11 to 14 years (median age: 12.5 years), participated in the study. Students engaged in an environmental science investigation, working in small groups to determine the cause of illness in farmed tilapia fish.

Each learning session lasted 90 minutes, with students interacting with the platform using laptops equipped with built-in webcams. The camera, positioned on an inclined support bar, recorded video at 30 frames per second (FPS), resulting in approximately 1,200,000 image frames. Given the sensitive nature of this dataset, original input images cannot be publicly displayed. Instead, only the seed images used for de-identification and the corresponding de-identified outputs will be included in our visualizations.

DAiSEE Dataset: To complement the classroom dataset, we also evaluate our method on DAiSEE (Dataset for Affective States in E-Environments), a publicly available dataset designed for automatic affect recognition in online learning environments. The dataset contains facial expression recordings of adult students (undergraduate, postgraduate, and graduate levels) engaging in e-learning activities. DAiSEE consists of 9,068 video snippets collected from 112 students (32 female, 80 male), covering a total of 2,723,882 frames (≈ 25 hours of recorded data). Each video is labeled for Engagement, Boredom,

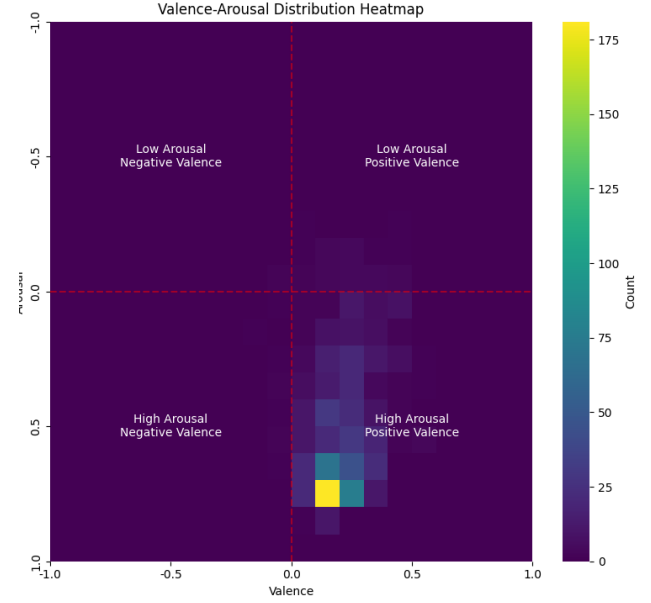


Figure 2: Valence Arousal Heatmap for all seeds

Confusion, and Frustration, with each state categorized into four levels (Very Low, Low, High, and Very High).

While both datasets allow for a rigorous evaluation of our method, DAiSEE enables a full qualitative analysis by allowing visualization of input, seed, and de-identified images, whereas in the classroom dataset, only the seed and de-identified images can be displayed due to privacy constraints.

4.2 Results and Analysis

We evaluate the effectiveness of our privacy-preserving face de-identification approach through a series of quantitative and qualitative assessments. Our evaluation focuses on three key aspects: expression retention, biometric identity suppression, and perceptual similarity. Additionally, we assess the impact of de-identification on microexpression retention (FAU loss) and verify that GAN-based transformations do not introduce significant visual distortions.

Emotion Retention Rate: To quantify how well the de-identified images retain emotional expressions, we compare the valence-arousal values of the original and de-identified images using HSE-motion [28] and EMONet [32], two state-of-the-art emotion recognition models. Our results (Table 1) demonstrate that our method preserves emotional expressions with minimal discrepancy, as reflected in the low RMSE values across both datasets.

A paired t-test confirms that our valence-arousal guided seed selection significantly reduces RMSE compared to random seed selection ($p < 0.001$), validating its effectiveness in preserving affective expressions (Table 1).

To further validate the impact of valence-arousal guided seed selection, we compare it against a baseline where the seed image is chosen randomly. The baseline model resulted in significantly higher RMSE values (valence: 0.35, arousal: 0.38 in the classroom dataset; valence: 0.33, arousal: 0.37 in DAiSEE). The improvement in emotion retention demonstrates that our guided seed selection

Dataset	RMSE ↓		Baseline (Random Seed) RMSE ↓	
	Valence	Arousal	Valence	Arousal
Classroom Dataset	0.19 ± 0.04	0.22 ± 0.03	0.35 ± 0.05	0.38 ± 0.04
DAiSEE Dataset	0.21 ± 0.03	0.24 ± 0.02	0.33 ± 0.05	0.37 ± 0.03

Table 1: Results of RMSE values for valence and arousal between our method and baseline random seed selection. Mean ± standard deviation over the images in the datasets. Lower values indicate better emotion retention.

Dataset	Mean Abs. Diff ↓		Baseline (Random Seed) Diff. ↓	
	Valence	Arousal	Valence	Arousal
Classroom Dataset	0.05 ± 0.01	0.07 ± 0.02	0.19 ± 0.05	0.22 ± 0.04
DAiSEE Dataset	0.06 ± 0.02	0.08 ± 0.03	0.18 ± 0.04	0.21 ± 0.05

Table 2: Mean absolute difference in valence-arousal values between input and selected seed. Mean ± standard deviation over the images in the datasets. Lower values indicate better expression alignment.

significantly outperforms naive random selection, ensuring that expressions are preserved in a more natural and structured manner.

Since our methodology selects seeds with valence-arousal values similar to the input image, we analyze how close these values actually are compared to random seed selection. A Wilcoxon signed-rank test confirms that the difference in valence-arousal values between input and selected seeds is significantly smaller for our method than for random selection ($p < 0.001$, Table 2).

The valence-arousal distribution heatmap (Figure 2) illustrates the emotional range covered by our precomputed synthetic seed images. The heatmap shows that the majority of seed images fall within the neutral to positive valence range, with a concentration in the high-arousal positive valence region (e.g., smiling, engagement, delight). In contrast, expressions corresponding to negative valence (e.g., sadness, frustration, disengagement) are sparsely represented or entirely missing.

This distribution reflects a key limitation: while our method effectively retains and transfers neutral to positive expressions, it cannot fully reconstruct the entire spectrum of human emotions due to the absence of appropriate seed images in certain VA regions. This limitation is particularly relevant for affective computing in education, where studies show that students spend over 70% of their time in engaged or flow states, which align with the neutral-to-positive regions of the VA space. However, for rare cases where students experience frustration, boredom, or confusion (negative valence states), our method may not accurately replicate these expressions due to the lack of corresponding seed images.

Thus, while our valence-arousal guided seed selection ensures effective expression retention within the available VA space, the absence of seeds covering low-valence emotions restricts its ability to reconstruct certain affective states.

Microexpression Retention (FAU Loss Analysis): While valence-arousal analysis confirms macro-level expression retention, it does not account for subtle microexpressions, such as slight eyebrow movements or lip adjustments. To address this, we measure Facial

Action Unit (FAU) loss, which quantifies how well fine-grained muscle movements are preserved after de-identification. Using Affective SDK (iMotions), we compute FAU loss across both datasets. As shown in Table 3, our method achieves low FAU loss values (0.15 for Classroom and 0.18 for DAiSEE), confirming that fine-grained facial muscle movements remain consistent post-de-identification. However, without established benchmarks for FAU loss in de-identification contexts, it’s challenging to directly assess the significance of these values. To provide a relative perspective, we compared our results to a baseline method employing random seed selection, which yielded higher FAU loss values of 0.27 for the Classroom dataset and 0.29 for DAiSEE. This comparison suggests that our method offers a substantial improvement in microexpression retention, effectively preserving fine-grained facial muscle movements.

Dataset	FAU Loss (L2 Distance) ↓
Classroom Dataset	0.15 ± 0.03
DAiSEE Dataset	0.18 ± 0.02

Table 3: FAU loss values (L2 distance) measuring microexpression retention. Lower values indicate better preservation of fine-grained muscle movements.

Privacy Preservation and Identity Suppression: A fundamental requirement of de-identification is ensuring that biometric identity is effectively removed. To measure privacy preservation, we compute cosine similarity between the FaceNet embeddings [29] of the original and de-identified images. The privacy preservation score for the classroom dataset was 89.97%, while for DAiSEE, it was 89.94% (Table 4), confirming that the de-identified outputs have almost no traceable biometric identity linked to the original subjects.

To further validate privacy, we compute the re-identification rate, which represents the percentage of cases where a de-identified image can still be matched to its original identity. As shown in Table 4, our results indicate a near-zero re-identification rate, with only 0.015% of cases in the classroom dataset and 0.02% in DAiSEE being re-identified. These findings demonstrate that our approach

effectively anonymizes faces while maintaining emotional expressions.

Dataset	Our Method	
	Privacy Score $\uparrow$	Re-ID Rate (%) $\downarrow$
Classroom	89.97% $\pm$ 0.01%	0.015%
DAiSEE	89.94% $\pm$ 0.02%	0.02%

Table 4: Results of privacy scores (Cosine Similarity) and re-identification rates for our method.

Perceptual Similarity and Non-Biometric Feature Retention: To ensure that non-biometric facial characteristics (e.g., structure, skin texture) remain consistent after de-identification, we compute the Learned Perceptual Image Patch Similarity (LPIPS) score [37]. A higher LPIPS score suggests strong similarity between the original and de-identified faces in terms of structural and non-identifiable features.

As shown in Table 5, our method achieved an LPIPS score of 0.87 in the classroom dataset and 0.85 in DAiSEE, indicating that de-identified faces maintain general face structure while successfully anonymizing identifiable details.

Evaluating GAN-Based Distortion (FID Score): Since GAN-based de-identification techniques can introduce artifacts, we evaluate whether StyleGAN inversion and style mixing distort facial structures by computing the Fréchet Inception Distance (FID) [30] between the original and de-identified images. Lower FID values indicate better visual quality and minimal distortions. Our method achieved an FID score of 12.3 in the classroom dataset and 14.1 in DAiSEE, suggesting that StyleGAN-based de-identification does not introduce significant artifacts or distortions in the final images. To contextualize these results, we refer to established benchmarks in the literature: A study comparing CycleGAN and StyleGAN reported FID scores of 71.8 and 21.7, respectively, for high-resolution face generation tasks. Another comparative evaluation of StyleGAN and DCGAN found that StyleGAN produced superior quality images with lower FID scores compared to DCGAN [3].

Given these benchmarks, our FID scores of 12.3 and 14.1 suggest that our de-identified images are of high visual quality, comparable to or exceeding the performance of leading generative models like StyleGAN. This indicates that our de-identification process maintains the realism and quality of facial images while effectively anonymizing them.

Qualitative Visualization of De-Identification: In addition to quantitative evaluations, we provide qualitative comparisons of input, seed, and de-identified faces. Due to privacy restrictions in the classroom dataset, we only display seed images and de-identified

Dataset	LPIPS $\uparrow$	FID $\downarrow$
Classroom Dataset	0.87 $\pm$ 0.02	12.3 $\pm$ 1.5
DAiSEE Dataset	0.85 $\pm$ 0.03	14.1 $\pm$ 1.8

Table 5: Results of LPIPS and FID scores for evaluating perceptual similarity and GAN-based distortions. LPIPS $\uparrow$ (Higher = More Similar Non-Biometric Features. FID $\downarrow$ (Lower = Less GAN Distortion)).

Process	Time Per Frame (ms) $\downarrow$
GAN Inversion	114 ms
Valence-Arousal Seed Selection	28 ms
Traditional Blurring	4 ms

Table 6: Computational efficiency breakdown for different components of our de-identification pipeline.

outputs, whereas for DAiSEE, we present side-by-side visualizations of input images, seed images, and de-identified outputs. These comparisons confirm that facial expressions are well-preserved while identity-related features are suppressed, reinforcing our quantitative findings.

Our approach remains computationally feasible for real-world deployment. As shown in Table 6, inference time for de-identification is 0.17 seconds per frame on an NVIDIA GeForce RTX 2080 Ti GPU, making it suitable for batch processing of recorded video data. While GAN inversion introduces some computational overhead, the precomputed seed selection mechanism ensures faster execution compared to on-the-fly GAN-based anonymization methods.

Manual Verification of Expression Consistency: While our quantitative metrics confirm strong expression retention, we further validate our approach through a manual annotation study to assess perceived expression consistency. To further validate our automated expression retention metrics, we conducted a manual assessment on a randomly selected subset of 500 de-identified images across both datasets. Two independent annotators assessed whether the expressions in the de-identified images were perceptually consistent with the original input. Our results show that in 92.4% of cases, annotators agreed that the expressions remained visually consistent. To measure inter-rater reliability, we computed Cohen’s Kappa ($\kappa = 0.84$), indicating strong agreement between annotators. However, in 7.6% of cases, subtle variations in expression intensity were noted, particularly for high-arousal emotions (e.g. delight). This aligns with our FAU loss results, indicating that while macro-level expressions are well-preserved, fine-grained microexpressions can sometimes exhibit minor discrepancies.

4.3 Discussion

Evaluation of SOTA Models with Our Pipeline: To assess the generalizability of our method, we applied valence-arousal guided seed selection to state-of-the-art (SOTA) de-identification models, including CIAGAN, DeepPrivacy, and k-Same-Net [8, 19]. These models effectively suppress biometric identity, but they do not explicitly retain fine-grained emotional expressions, making them less suitable for affect-sensitive AI applications. Our results demonstrate that integrating our pipeline with CIAGAN reduced valence-arousal RMSE from 0.34–0.39 to 0.21–0.24, comparable to the performance of our StyleGAN-based de-identification approach. Similarly, FAU loss improved from 0.22 to 0.16 when using DeepPrivacy, confirming better microexpression retention.

While expression retention improved across these models, privacy performance remained dependent on the underlying de-identification mechanism. CIAGAN and DeepPrivacy achieved privacy scores of 89.5% and 88.9%, respectively, whereas our StyleGAN-based model

attained 89.97%. This highlights an important distinction: our contribution lies in expression-preserving de-identification, while overall privacy performance is contingent on the base identity suppression technique. These findings confirm that our pipeline is model-agnostic and enhances affect retention across different de-identification frameworks, rather than competing directly with identity suppression techniques.

Why Baseline Comparisons Are Not Meaningful: Unlike traditional de-identification methods, which prioritize identity suppression, our work ensures that de-identified faces maintain the input’s original emotional state. This fundamental difference makes direct baseline comparisons inappropriate. SOTA models such as DeepPrivacy and k-Same-Net were not designed for emotion preservation, so evaluating them as baselines would only measure how well identity is removed—not how well emotions are retained.

Additionally, comparing our approach to conventional identity suppression models would fail to account for cases where expression mismatches occur. For example, if a method assigns a randomly selected face to a de-identified identity, the resulting expression could become inconsistent with the original, distorting affect recognition outputs. In contrast, our seed selection mechanism explicitly aligns the synthetic face’s valence-arousal values with the input, ensuring affective consistency. Therefore, comparing against traditional de-identification baselines would conflate privacy preservation with expression retention—two fundamentally distinct objectives.

Impact of AI-Generated vs. Real (Celebrity) Seed Images: An important factor influencing privacy performance is the choice of seed images. Some methods use AI-generated synthetic faces, as in our approach, while others rely on real faces from celebrity datasets. Our findings show that using AI-generated seeds leads to stronger privacy guarantees, with a privacy score of 89.97% and a re-identification rate of only 0.015%. In contrast, when celebrity faces were used, privacy scores dropped to 88.4%, with a re-identification rate of 0.09%.

This discrepancy arises because real faces inherently retain identifiable biometric patterns, increasing the risk of partial feature overlap between the input and anonymized output. Even after de-identification, real-world datasets contain residual features that can potentially be traced back to an individual. In contrast, AI-generated faces are fully disentangled from real identities, ensuring that no identifiable information is retained.

Additionally, using real faces can result in expression mismatches if the seed image’s affective characteristics do not align with the input face. For example, if a student displaying mild confusion (valence: -0.2, arousal: 0.4) is mapped to a celebrity face with a neutral expression (valence: 0.0, arousal: 0.1), the resulting face may blend these expressions incorrectly, leading to unreliable affective inferences. Our AI-generated seed selection mechanism eliminates this issue by ensuring that the selected seed face precisely matches the original valence-arousal values, preventing emotion distortions.

Ethical and Legal Compliance Considerations: Recital 26 of the General Data Protection Regulation (GDPR) makes clear that data protection principles apply to any information concerning an identified or identifiable natural person, including data that has undergone pseudonymisation (i.e. where re-identification may be possible via additional information). The protection regime does not extend

to data that has been rendered anonymous in such a manner that the data subject is not or no longer identifiable, considering all means reasonably likely to be used (e.g., via cost, time, or technological advances). In designing our de-identification framework, we aim for irreversibility and robust resistance against re-identification, thereby approaching the standard of anonymity under GDPR. However, we acknowledge that perfect anonymity is difficult in practice, especially given future advances in re-identification techniques.

In the U.S. educational context, our approach similarly addresses the requirements under Family Educational Rights and Privacy Act (FERPA): by ensuring that de-identified facial data cannot feasibly be linked back to individual identities, we support lawful data sharing and secondary analyses under FERPA’s protections of student PII (Personally identifiable information). Together, these considerations show that our method is not merely a technical mechanism but is designed with ethical and regulatory sensitivity for deployment in real educational settings.

Broader Implications: Data Sharing & Real-World Classroom Applications: Beyond privacy preservation, our approach offers two significant real-world benefits. First, it enables the creation of privacy-preserving emotion recognition datasets, allowing researchers to publicly share affective computing datasets while safeguarding individual identities. This is especially critical for studying children’s emotional responses, where ethical and regulatory constraints often prevent open data sharing. By making de-identified datasets publicly available, our method can enhance the robustness of emotion recognition models, improving AI systems designed for education, healthcare, and affective computing research.

Second, our method is well-suited for real-time deployment in educational settings. In Intelligent Tutoring Systems (ITS), facial expressions are frequently used to track engagement, adapt instructional strategies, and provide real-time feedback. However, privacy concerns have hindered the widespread use of facial analysis in classrooms. Our approach resolves this issue by de-identifying student faces in real-time before sending them to a central server for emotion analysis, ensuring that learning analytics can be performed without exposing personal identity information. Furthermore, our method is resistant to reverse engineering, meaning that the original face cannot be reconstructed from the de-identified image. This property makes it an ideal solution for privacy-compliant affective computing applications in classrooms, where facial expressions can be analyzed without compromising student anonymity.

5 Conclusion

This paper presents a privacy-preserving face de-identification framework designed for affect-sensitive AI applications in educational settings. By leveraging valence-arousal guided identity replacement within the StyleGAN latent space, we achieve a balance between privacy and utility, making our method particularly suited for learning analytics and AI-driven tutoring systems. Our experimental evaluations on both a real-world classroom dataset and the DAiSEE dataset demonstrate that our framework achieves 89.97% and 89.94% privacy preservation accuracy, respectively, while maintaining high expression fidelity with a 92.4% manual verification agreement and low FAU loss (0.15–0.18).

While our method achieves strong privacy preservation and expression retention, several areas warrant further exploration. One key direction is improving temporal consistency in video-based de-identification, as processing frames independently can introduce inconsistencies. Future work could explore spatiotemporal GANs or diffusion models with recurrent constraints to ensure smooth transitions in expression dynamics. Another avenue is enhancing adaptive latent space manipulation, where self-supervised learning or disentangled representation learning could refine latent codes for more precise expression preservation beyond valence-arousal matching.

Acknowledgment

The National Science Foundation AI Institute Grant #DRL-2112635 supported this research. Any opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Celestine E Akpanoko, Gautam Biswas, et al. 2024. The interplay of affective states and cognitive processes in an open-ended learning environment: a case study. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024*, pp. 873–880. International Society of the Learning Sciences.
- Celestine E Akpanoko, Grayson Cordell, Gautam Biswas, et al. 2024. Investigating the relations between students' affective states and the coherence in their activities in open-ended learning environments. In *Proceedings of the 17th International Conference on Educational Data Mining*, 511–517.
- M Archana, Ashwin Antony Joseph, Alex V Ajiith, Alan N Koshy, and S Saritha. 2024. Stylegan and dcgan for face generation: a comparative evaluation. In *Proceedings of Eighth International Conference on Information System Design and Intelligent Applications*. Springer, 517–524.
- T. S. Ashwin and Ram Mohana Reddy Guddeti. 2020. Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. en. *User Modeling and User-Adapted Interaction*, 30, 5, (Nov. 2020), 759–801. doi:10.1007/s11257-019-09254-3.
- TS Ashwin and Ram Mohana Reddy Guddeti. 2018. Unobtrusive students' engagement analysis in computer science laboratory using deep learning techniques. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 436–440.
- TS Ashwin and Ramkumar Rajendran. 2023. Preserving privacy of face and facial expression in computer vision data collected in learning environments. In *International Conference on Artificial Intelligence in Education*. Springer, 561–567.
- Allison Macey Banzon, Jonathan Beever, and Michelle Taub. 2023. Facial expression recognition in classrooms: ethical considerations and proposed guidelines for affect detection in educational settings. *IEEE Transactions on Affective Computing*, 15, 1, 93–104.
- Jingyi Cao, Xiangyi Chen, Bo Liu, Ming Ding, Rong Xie, Li Song, Zhu Li, and Wenjun Zhang. 2024. Face de-identification: state-of-the-art methods and comparative studies. *arXiv preprint arXiv:2411.09863*.
- Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction*, 22, 2, 145–157.
- Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.
- Hang Du, Hailin Shi, Dan Zeng, Xiao-Ping Zhang, and Tao Mei. 2022. The elements of end-to-end deep face recognition: a survey of recent advances. *ACM computing surveys (CSUR)*, 54, 10s, 1–42.
- Joyce Fonteles et al. 2024. A first step in using machine learning methods to enhance interaction analysis for embodied learning environments. In *Artificial Intelligence in Education*. Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, (Eds.) Springer Nature Switzerland, Cham, 3–16. ISBN: 978-3-031-64299-9.
- Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. 2022. Stylegan-human: a data-centric odyssey of human generation. In *European Conference on Computer Vision*. Springer, 1–19.
- Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. Daisee: towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: a generative adversarial network for face anonymization. In *International symposium on visual computing*. Springer, 565–578.
- Maryam Imani and Gholam Ali Montazer. 2019. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147, 102423.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.
- Wasiq Khan, Luke Topham, Umar Khayam, Sandra Ortega-Martorell, Panter Heather, Darren Ansell, Dhiya Al-Jumeily, and Abir Hussain. 2024. Person de-identification: a comprehensive review of methods, datasets, applications, and ethical aspects along-with new dimensions. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.
- Dimitrios Kollias, Attila Schulc, Elnar Hajiyeve, and Stefanos Zafeiriou. 2021. Analysing affective behavior in the first abaw 2020 competition. *IEEE Transactions on Affective Computing*.
- Jhon Lopez, Carlos Hinojosa, Henry Arguello, and Bernard Ghanem. 2024. Privacy-preserving optics for enhancing protection in face de-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12120–12129.
- Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. 2019. Automatic Analysis of Facial Actions: A Survey. *IEEE Transactions on Affective Computing*, 10, 3, (July 2019), 325–347. doi:10.1109/TAFFC.2017.2731763.
- Bradford W Mott, Robert G Taylor, Seung Y Lee, Jonathan P Rowe, Asmalina Saleh, Krista D Glazewski, Cindy E Hmelo-Silver, and James C Lester. 2019. Designing and developing interactive narratives for collaborative problem-based learning. In *International Conference on Interactive Digital Storytelling*. Springer, 86–100.
- Marius Muja and David G Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications*, 331–340.
- A Pena-Molina and MM Larrondo-Petrie. 2023. Data privacy and security in online laboratory management systems. In *INTED2023 Proceedings*. IATED, 6459–6465.
- James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39, 6, 1161.
- Ashwin TS and Gautam Biswas. 2024. Relating students cognitive processes and learner-centered emotions: an advanced deep learning approach. In *Proceedings of the International Conference on Multimodal Interaction (ICMI '24)*. ACM, San Jose, Costa Rica, (Nov. 2024).
- Andrey Savchenko. 2023. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *International Conference on Machine Learning*. PMLR, 30119–30129.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: a unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, (June 2015), 815–823. doi:10.1109/cvpr.2015.7298682.
- Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>. Version 0.3.0. (Aug. 2020).
- Adrian Suwała, Bartosz Wójcik, Magdalena Proszewska, Jacek Tabor, Przemysław Spurek, and Marek Śmieja. 2024. Face identity-aware disentanglement in stylegan. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5222–5231.
- Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3, 1, 42–50.
- Ashwin TS and Ram Mohana Reddy Guddeti. 2020. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and information technologies*, 25, 2, 1387–1415.
- Zhihong Xu, Yingying Zhao, Bingsheng Zhang, Jeffrey Liew, and Ashlynn Kogut. 2023. A meta-analysis of the efficacy of self-regulated learning interventions on academic achievement in online and blended environments in k-12 and higher education. *Behaviour & Information Technology*, 42, 16, 2911–2931.
- Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. Facial-video-based physiological signal measurement: recent advances and affective applications. *IEEE Signal Processing Magazine*, 38, 6, 50–58.
- Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23, 10, 1499–1503.
- Richard Zhang, Phillip Isola, and Alexei A Efros. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.