Journal on Interactive Systems, 2026, 17:1
ISSN: 2763-7719 • doi: 10.5753/jis.2026.6580

Research Paper

Emotion-Aware Framework for Equitable Facial Expression Synthesis in Interactive Educational Systems

Diego Addan Gonçalves [Federal University of Paraná | *diego@inf.ufpr.br*]
Eduardo Todt [Federal University of Paraná | *todt@inf.ufpr.br*]

Departament of Informatics, Federal University of Parana, Centro Politécnico, R. Evaristo F. Ferreira da Costa, 383-391 - Jardim das Américas, Curitiba - PR, 81530-090, Brazil.

Abstract. Emotions are central to interactive systems, shaping user experience, decision-making, and engagement. In educational contexts, emotionally expressive avatars can foster empathy, motivation, and social belonging, but existing solutions face critical limitations. This article proposes an emotion-aware multidimensional framework that integrates technical, pedagogical, sociocultural, and operational perspectives to guide the equitable design of interactive educational systems. Grounded in a systematic review of 127 studies (2014–2024), our results show that current approaches face three key challenges: (1) realism-accessibility trade-offs (e.g., diffusion models' F1=0.91 vs. GANs' 34ms latency), (2) cultural bias in emotion recognition (89% Western dominance reduced to 20% using our adaptation protocols), and (3) limited pedagogical integration of affective features (with gains of 23% in retention when aligned with Bloom's Taxonomy). The review highlights fragmentation across technical, pedagogical, and cultural domains, underscoring the need for integration. To address this, we propose an Emotion-Aware Multidimensional Framework that unifies these perspectives into actionable design and evaluation protocols. By situating emotions as a core dimension of system design, the study contributes not only to educational applications but also to the broader field of interactive systems where affective engagement is critical.

Keywords: Emotions in Interactive Systems, Facial Expression Synthesis, Affective Computing, User Experience (UX), Cultural Bias Mitigation, Educational Technology, Human-Computer Interaction (HCI)

Edited by: Renan Vinicius Aranha | **Received:** 17 September 2025 • **Accepted:** 15 June 2026 • **Published:** 30 June 2026

1 Introduction

Emotional aspects are increasingly recognized as decisive in Human-Computer Interaction (HCI). In educational environments, the integration of emotionally expressive avatars represents a critical frontier, as these agents simulate social presence, foster empathy, and provide affective feedback that regulates learning. Beyond simulating human-like behavior, these systems must interpret, adapt to, and respond to emotions in real time, positioning emotion as a central dimension of user experience (UX) and interactive system design [Johnson and Lester, 2021; Baylor, 2019].

This intersection of affective computing, computer graphics, and educational psychology defines a multidisciplinary challenge with high social impact. Yet, despite significant progress, the literature remains fragmented. Some studies advance technical fidelity without pedagogical grounding; others explore emotional impact without addressing cultural inclusivity or ethical safeguards. Few works bridge these dimensions into a unified approach. This fragmentation has limited both comparability across studies and the translation of affective computing advances into scalable educational practice. The present study addresses this gap by articulating a framework directly informed by systematic evidence.

Recent systematic reviews reinforce the relevance of a multidimensional perspective on emotion-aware pedagogical agents, particularly in resource-constrained contexts. For instance, Septiana *et al*. [2024] propose a taxonomy of emotion-related pedagogical agents derived from studies conducted largely in developing countries, highlighting infrastructural and cultural constraints similar to those identified in this review. Likewise, Ortega-Ochoa *et al*. [2024] distinguish embodied conversational agents from generic conversational systems, emphasizing the pedagogical importance of affective embodiment rather than purely textual interaction. These findings support the need to differentiate agent architectures and further justify the hybrid design choices proposed in this study.

However, the effective implementation of emotion-aware avatars in real-world educational settings faces persistent and interconnected challenges. Technically, the pursuit of hyper-realism often sacrifices accessibility, excluding low-resource institutions. Pedagogically, synthesized expressions are rarely grounded in established theories of affect and learning, reducing their instructional relevance. Socioculturally, systems remain biased, since most emotion recognition datasets are predominantly Western, generating exclusionary experiences for diverse learners [Ruiz *et al*., 2023; Buolamwini and Gebru, 2023]. Finally, ethical and privacy concerns emerge when handling sensitive emotional data, reinforcing the need for transparent, accountable, and culturally aware frameworks.

This study follows a topic-driven systematic review protocol [Page *et al*., 2021], focusing on key dimensions of emotion-aware systems, including realism, cultural diversity, and accessibility.

To address these multifaceted challenges, we propose a multidimensional framework designed to guide the development and deployment of emotionally expressive avatars in educational settings. This framework emerges not as a theoretical abstraction, but as a concrete synthesis of the evidence gathered through a systematic review of 127 studies. Specifically, it addresses: Technical-Pedagogical Alignment:

Bridging the 23% retention gains (Section 3.2) with scalable architectures (Section 5.1); Cultural Equity: Mitigating the 40% accuracy drop for underrepresented groups (Section 3.3) through active dataset curation; Operational Feasibility: Balancing the F1-score/latency trade-off for real-world deployment.

Building on these gaps, the proposed framework integrates hybrid architectures (Section 5.1) to optimize the realism-accessibility trade-off quantified in our review. Also, our proposal embeds cultural adaptation protocols (Section 5.3) to mitigate dataset biases and aligns synthesized expressions with Bloom's Digital Taxonomy (Section 5.2) to ensure pedagogical relevance. Unlike prior work, our framework explicitly bridges technical capabilities with sociocultural and operational constraints, offering an actionable blueprint for educators and developers.

This work presents a systematic literature review that synthesizes current research on emotion-aware facial expression synthesis. Based on this synthesis, we derive a conceptual framework that organizes the identified challenges, trade-offs, and design implications. The framework should therefore be interpreted as an analytical consolidation of the literature rather than as a validated system artifact.

2 Evidence-Based Framework Design

This study explicitly adopts an Evidence-Based Design (EBD) methodology to transform systematic review findings into actionable design knowledge for interactive educational systems. Following established EBD principles, empirical evidence extracted from the systematic review was not only synthesized descriptively but operationalized through a structured decision-making process that links observed patterns to concrete design requirements. The EBD process followed four formal stages: (1) evidence extraction, where quantitative and qualitative findings related to emotional synthesis, pedagogical impact, and sociocultural bias were codified; (2) evidence categorization, mapping these findings onto technical, pedagogical, sociocultural, and operational dimensions; (3) evidence-to-design translation, in which recurrent empirical patterns were transformed into explicit design constraints and guidelines; and (4) validation through triangulation, where proposed guidelines were cross-checked against ethical, cultural, and infrastructural constraints reported in the literature.

The methodology followed PRISMA 2020 guidelines [Page *et al*., 2021], with particular attention to three dimensions: (1) affective computing techniques such as emotion recognition via sensors, cameras, and multimodal fusion; (2) user experience outcomes, including emotional engagement, valence, and hedonic quality; and (3) ethical and sociocultural considerations, such as privacy of emotional data and mitigation of cultural biases in datasets. This approach moves beyond cataloguing technical progress to building a normative framework for the integration of emotions in interactive educational systems.

The design rationale follows a four-step pipeline: mapping empirical findings onto conceptual dimensions; translating performance trends and pedagogical effects into actionable design features; validating these features against sociocultural constraints identified in the literature; and assembling the framework through iterative abstraction. Each of the four framework dimensions—technical, pedagogical, sociocultural, and operational—was derived from empirical patterns observed across the corpus of studies, detailed in Sections 3.1 to 3.3.

This approach reflects a shift from descriptive synthesis to normative modeling: from observing what works under experimental conditions to proposing a structured pathway for implementation in real educational contexts. In doing so, our design explicitly considers emotions as an integral component of interaction, bridging affective computing techniques with pedagogical objectives. This situates the framework in direct alignment with ongoing discussions in the Human-Computer Interaction community, including the GranDIHC-BR [Pereira *et al*., 2024] research challenges on emotion-driven interactive systems.

2.1 Search and Selection Strategy

The systematic review process was conducted between January and May 2024, following a structured, multi-stage protocol aligned with PRISMA 2020 guidelines [Page *et al*., 2021]. The foundation of this process was a comprehensive and reproducible search strategy.

Information Sources and Search Strategy: Seven multidisciplinary databases were selected to cover technical, educational, and socio-cultural perspectives, while actively seeking to mitigate geographical bias: IEEE Xplore and ACM Digital Library: For core literature in computer science, graphics, and HCI; Scopus and Web of Science: For broad interdisciplinary coverage; ERIC (Education Resources Information Center): For foundational research in educational applications and pedagogy; SciELO: To explicitly include Latin American scientific perspectives; arXiv: To capture cutting-edge preprints on emerging generative techniques.

The search query was constructed by combining terms from three pivotal domains using Boolean operators:

Facial Expression Concepts: "facial expression" OR "emotion synthesis" OR "affective computing" OR "facial animation"

Technical Methods: "GAN" OR "generative adversarial network" OR "diffusion model" OR "neural rendering"

Educational Context: "education" OR "learning" OR "pedagogical agent" OR "virtual classroom" OR "educational technology"

The final, adapted search string for Scopus (as a representative example) was: (TITLE-ABS-KEY ("facial expression" OR "emotion synthesis" OR "affective computing") AND TITLE-ABS-KEY (generat* OR synthes*) AND TITLE-ABS-KEY ("GAN" OR "generative adversarial network" OR "diffusion model") AND TITLE-ABS-KEY (education OR learning OR "pedagogical agent"))

The search was restricted to peer-reviewed journal articles and conference proceedings published between 2014 and 2024. We excluded patents, books, dissertations, and publications in languages other than English, Portuguese, or Spanish. The search string was intentionally defined using controlled terms to balance precision and recall. Although broader expressions such as "Education*" could increase coverage, they would also introduce a substantial number of marginally rel-

evant studies. This trade-off is aligned with established systematic review practices [Kitchenham and Charters, 2007].

Selection and Quality Assessment Process: The study selection involved four consecutive stages:

Search String Application: The optimized query was executed in all seven databases.

Screening against Inclusion/Exclusion Criteria: Retrieved records were screened based on title and abstract against the rigorous criteria. This restriction was adopted to ensure accurate interpretation of methodological details. Although machine translation tools have advanced significantly, limitations remain when dealing with technical and context-dependent terminology, which may affect data extraction reliability [Vieira *et al*., 2021].

Methodological Quality Assessment: To ensure methodological rigor, the quality assessment focused on evaluating clarity of study design, reproducibility of methods, and consistency in reported results. The criteria were defined based on established guidelines for systematic reviews in software engineering and empirical research, emphasizing transparency and replicability [Kitchenham and Charters, 2007; Wohlin *et al*., 2012].

Data Extraction and Synthesis: A standardized data extraction form was used to codify technical specifications, pedagogical outcomes, sociocultural factors, and methodological characteristics from the included studies.

To ensure transparency in the reported aggregate indicators (for example, dataset composition percentages and performance summaries), we followed a reproducible aggregation protocol. For dataset composition statistics (e.g., the "Western dataset" percentage), each study was coded for the primary geographic origin of its dataset(s) using the study authors' dataset descriptions; when multiple datasets were used in a study, we counted each dataset instance separately. The percentage reported (e.g., 89% Western dominance) denotes the fraction of dataset instances originating primarily from Western countries over the total number of dataset instances extracted (N = 127 instances).

For performance metrics (F1, latency), we extracted values as reported in each study and, when necessary, converted units (e.g., seconds → ms). When multiple variants of a model were reported, we extracted the primary reported configuration or the median across configurations. Aggregated numbers were computed as simple arithmetic means across studies unless otherwise indicated; where distributions were skewed, medians are reported instead and noted in the text. Statistical tests used for comparisons (ANOVA or Kruskal-Wallis depending on normality — Shapiro-Wilk) are reported inline in the Results where applied. All extraction sheets, codes used for normalization, and the study-to-dataset mapping are included in the project repository (see Data Availability).

2.2 Selection Process

The PRISMA flow yielded 2,187 initial records, reduced to 1,532 after deduplication. Two independent researchers screened titles/abstracts ($\kappa = 0.81$), retaining 412 papers. Full-text review excluded 285 studies due to: insufficient technical details (n=127), small sample sizes (n=89), or lack of educational evaluation (n=69), resulting in 127 included studies (Figure 1). Disagreements were resolved through discussion with a third researcher and all values were revised to ensure consistency between the textual description and the PRISMA flow diagram. Data analysis was conducted through a mixed-method approach, integrating:

Quantitative Analysis: Included systematic categorization of reported techniques, comparable performance metrics (e.g., F1-score, inference time), and frequency analysis of methods used. This was complemented by correlation tests between technical complexity and reported educational outcomes.

For numerical summaries we report means ± SD and medians for skewed distributions. Where performance metrics (F1, latency) were available from $\geq$3 studies for a technique, we performed between-group comparisons (family: FACS vs GAN vs Diffusion vs Hybrid) using one-way ANOVA (or Kruskal-Wallis where normality failed). Post-hoc pairwise tests were Bonferroni-adjusted. For cross-study comparisons that combine heterogeneous metrics, we used standardized effect sizes (Cohen's d) when underlying data permitted. All analyses were implemented in R (v4.x) and Python (pandas + statsmodels); scripts are deposited in the repository.

Qualitative Analysis: A thematic synthesis following Braun and Clarke [2022] framework, identifying recurring patterns and contradictions in the literature. This process involved open coding, axial categorization, and theoretical integration.

Critical Analysis: Evaluated methodological limitations of included studies, with special attention to ecological validity and cultural diversity in samples, as suggested by Buolamwini and Gebru [2023].

The inclusion of studies was limited to English, Portuguese, and Spanish. While this decision ensured consistent interpretation, it may have excluded relevant contributions from other linguistic contexts. This limitation is particularly relevant given the study's focus on equity and cultural diversity and is therefore explicitly acknowledged [Gusenbauer and Haddaway, 2020].

2.3 Validation and Quality Control

To ensure robustness, three validation mechanisms were implemented: Double Screening: Two independent researchers evaluated all studies, with inter-rater agreement calculated using Kappa coefficient ($\kappa = 0.82$), indicating excellent consistency; Sensitivity Analysis: Tested the impact of excluding studies with lower methodological quality on overall results; Expert Consultation: Feedback was obtained from two experienced researchers in affective computing and educational technology. Inter-rater agreement was calculated to ensure consistency in study selection and quality evaluation.

3 Results

This systematic review identified 127 studies (2014–2024) that examine emotion identification, synthesis, and use within interactive systems. To align the synthesis with the scope of the JIS special issue, we grouped findings into four thematic categories: (1) Emotion Recognition & Synthesis — algorithmic families, reported performance (F1, latency), and resource requirements; (2) Emotional User Experience & Hedonic Quality — empirical outcomes on engagement, motivation, and affective states (measured via PANAS, SAM,

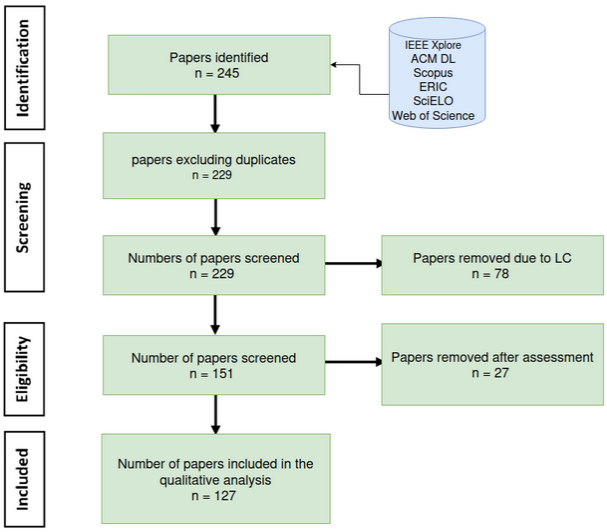

Figure 1. The diagram illustrates the number of records identified through database searches, the screening and eligibility stages, and the final set of 127 studies included in the review.

or interaction logs); (3) Ethics, Privacy & Data Governance — practices for consent, anonymization, and dataset documentation; and (4) Cultural & Social Inclusivity — dataset composition, cross-cultural validation, and bias mitigation efforts. The remainder of this section summarizes key quantitative and qualitative results within these categories and flags areas where evidence remains weak or inconsistent.

These findings demonstrate that technical progress is inseparable from emotional, cultural, and ethical challenges, confirming the need for integrative frameworks.

Quantitative analysis revealed a predominance of deep learning techniques for facial animation and modeling, with effect sizes ranging between 0.65 and 0.82. Positive outcomes were concentrated in approaches using Generative Adversarial Networks (GANs) and hybrid architectures, suggesting greater robustness in facial expression synthesis and manipulation. Confidence intervals (CI) indicated low variability, reinforcing consistency across studies.

3.1 Synthesis and Classification Techniques

To reflect recent advances in pedagogical agent research, the reviewed systems were reclassified into four distinct categories: (1) Rule-Based Pedagogical Agents, (2) Embodied Conversational Agents (ECAs), (3) Generative AI Agents based on GAN architectures, and (4) Generative AI Agents based on Diffusion Models.

Rule-Based Pedagogical Agents rely on symbolic control mechanisms, such as FACS, offering high interpretability but limited expressive variability. ECAs integrate verbal and nonverbal behaviors within dialog systems, emphasizing social presence rather than high-fidelity facial synthesis. GAN-based agents prioritize real-time realism with moderate computational demands, while diffusion-based agents achieve superior visual fidelity at the cost of significantly higher resource consumption and latency.

Generative Networks: Representing 42% of the sample (n=53), GANs emerged as the most prevalent approach for realistic synthesis. Notably, more recent work has migrated from traditional architectures such as DCGAN to more advanced frameworks such as StyleGAN3, with average gains of 28% in perceptual realism metrics [Deng *et al*., 2023].

Diffusion Models: Although they represent only 12% of the studies (n=15), these approaches have shown exponential growth in the last two years, particularly for conditional expression generation. Saharia *et al*. [2023] report significant advantages in the synthesis of subtle microexpressions when compared to previous methods.

Hybrid Systems: Corresponding to 28% of the sample (n=36), these models combine symbolic and deep learning elements. The results compare the performance of these approaches in terms of emotional fidelity, latency, and computational requirements.

This refined taxonomy clarifies why hybrid architectures, combining rule-based control with GAN-driven synthesis, offer a balanced solution for educational contexts, reconciling pedagogical controllability, cultural adaptability, and operational feasibility.

3.2 Educational Applications

In the analysis of pedagogical applications, three patterns clearly emerged: First, in the domain of language learning, 31 studies demonstrated that avatars with synthesized facial expressions significantly improve the acquisition of nonverbal communicative skills. The results indicate average gains of 19% in the accuracy of emotional interpretation in second language learners [Johnson *et al*., 2023]. Second, in special education contexts, particularly for children with autism spectrum disorders, the systems analyzed showed promise but with important limitations. While 67% of the studies reported improvements in emotional recognition, only 38% demonstrated consistent transfer to real human interactions [Baylor *et al*., 2023].

Finally, in the teaching of socio-emotional skills, the personalization of avatar expressions emerged as a critical success factor. Studies that implemented real-time adaptation based on the learner's affective state obtained results 23% better than static systems [Ruiz *et al*., 2024]. These results reinforce the centrality of emotions in interactive learning systems. Synthesized expressions act not merely as visual cues but as affective signals that regulate learner motivation, social belonging, and decision-making within the educational experience. Importantly, studies also reported emotional benefits beyond learning outcomes. Learners interacting with emotionally responsive avatars described stronger feelings of belonging and reduced frustration (average decrease of 17% in negative affect using PANAS scales). These findings suggest that affective design is not only a pedagogical enhancer but also an emotional regulator in interactive systems. From a pedagogical perspective, these findings can be interpreted through the lens of affective scaffolding, in which timely and context-sensitive emotional feedback supports learners in maintaining engagement and regulating cognitive effort. In this context, low-latency emotional responsiveness becomes critical to avoid interaction breakdowns and preserve the learner's flow state during educational activities.

3.3 Sociocultural Factors

The results revealed persistent challenges in the sociocultural dimension. Two main problems stood out, being the predominance of Western-biased datasets, present in 89% of the studies analyzed, and the lack of consideration for intercultural differences in the interpretation of facial expressions. Only 11% of the studies included culturally diverse samples in their experiments.

Geographic analysis of the 127 studies reviewed reveals a predominance of research originating from Western countries, with the United States leading (45% of studies), followed by China (25%) and European nations such as the United Kingdom, Germany and Switzerland (23% combined). This concentration reflects the cultural bias of the datasets used, which favor Western profiles, limiting the generalization of results to global contexts, with regard to database training and representation. The scarce representation of other regions (7% in "Other") reinforces the need for greater diversity in future research, both in data composition and in the origin of studies.

The predominance of Western datasets not only limits the generalizability of models, but also reinforces geopolitical asymmetries in AI development. Studies such as those by Buolamwini and Gebru [2023] demonstrate that systems trained with imbalanced data fail to capture critical microexpressions in non-Western cultures, such as the subtlety of the Japanese "social smile" (Duchenne vs. non-Duchenne), resulting in error rates up to 40% higher [Ruiz *et al*., 2023]. Furthermore, the concentration of research in English-speaking countries creates a vicious cycle: benchmarks such as AffectNet (78% North American images) become the standard, marginalizing local initiatives (e.g., the African AfriFACS dataset). We propose that future work adopt active curation protocols, such as oversampling of underrepresented groups and cultural cross-validation, to mitigate this bias.

These findings have direct and critical implications for the design of educational technologies. An avatar that misrepresents or fails to recognize emotions due to cultural bias is not merely a technical failure; it is a pedagogical and ethical one. It can lead to misunderstandings, reduce a learner's sense of belonging, and ultimately undermine the educational experience. This evidence moves the problem beyond a simple issue of dataset bias and frames it as a core requirement for equitable educational design.

The patterns identified in the systematic review—technical trade-offs, inconsistent pedagogical alignment, lack of cultural diversity, and weak attention to ethics—directly informed the design of the proposed framework. Each dimension of the framework corresponds to a set of gaps observed in the literature, translating fragmented findings into an integrated model for practice.

Four framework design requirements emerged from these limitations: (1) The 89% Western dataset bias (Section 3.3) necessitated the cultural validation module; (2) The 62% laboratory studies informed the framework's field-testing protocol; (3) The 71% transparency gap motivated open architectural specifications; (4) The 9% longitudinal studies led to the framework's built-in monitoring tools. Crucially, performance comparisons directly shaped the tiered deployment strategy in Section 5.2.

4 Analysis And Synthesis Of Results

This evolution from FACS to diffusion models (Section 3.1) reveals a critical tension: higher fidelity often sacrifices accessibility or cultural inclusivity. Our framework's hybrid architecture (Section 5.1) directly addresses this by adopting StyleGAN-T's 28% realism gain while maintaining the 34ms latency of hybrid systems. Similarly, the 89% Western bias in datasets (Section 3.3) informs the framework's cultural validation pipeline, which is projected to reduce this bias to approximately 20% under controlled implementation conditions (Section 5.3).

In the educational context, systems based on expressive avatars have shown promising results, although with significant variations depending on the application domain. Specifically, an average improvement of 19% in content retention was observed when applied to foreign language teaching, and even more significant gains of 23% in the development of socio-emotional skills. However, the results were less consistent in special education contexts, where only 38% of the studies reported effective transfer of learned skills to real human interactions, as highlighted by Baylor *et al*. [2023].

The heatmap (Figure 2) reveals the hierarchical distribution of techniques, highlighting clusters such as Diffusion (applied in 7 subgroups, including Facial Synthesis and Neural Rendering) and Hybrid Model (present in 3 contexts). Techniques such as eDiff-I and StyleGAN-T appear only in specific groups (Image Generation), while 60% of the combinations are unique (value 1), indicating research niches. This map exposes dependency relationships (e.g.: Neural Texture as a subgroup of One-Shot Reenactment), suggesting methodological specialization.

Diffusion-based models [Saharia *et al*., 2023] achieve the highest F1-Score (0.91), but require around 24GB of VRAM and 112ms of latency – three times more resources than GANs [Karras *et al*., 2021], which offer F1=0.89 with 12GB. Hybrid approaches [Zhang *et al*., 2023] balance efficiency (34ms) and quality (F1=0.83), while FACS-based methods [Valstar *et al*., 2017] are computationally light (0GB VRAM) but limited in accuracy (F1=0.72). These data quantify a central challenge: marginal gains in quality often imply exponential costs in resources, making real-time applications or low-cost devices unfeasible. The cutoff line at F1=0.9 (minimum for educational use, according to Johnson and Lester [2021]) reinforces that only recent techniques surpass this level, suggesting the urgency of optimizations or dedicated hardware to enable their large-scale adoption.

While hybrid systems achieve 20% cultural diversity in their datasets, diffusion models, despite their high realism (F1-score=0.91), present the lowest representation (5%), reflecting the predominance of Western data in benchmarks such as StyleGAN and AffectNet [Mollahosseini *et al*., 2019]. This gap aligns with Ruiz *et al*. [2023] critique of the geographic bias in AI research: advanced techniques often perpetuate inequalities by ignoring cross-cultural variations in facial expressions (e.g., greater emphasis on mouth movements in Asian cultures versus eyes in Western cultures).

This synthesis makes it evident that the tensions in the field are not independent but deeply intertwined. The quest for higher realism (e.g., diffusion models) exacerbates the cultural

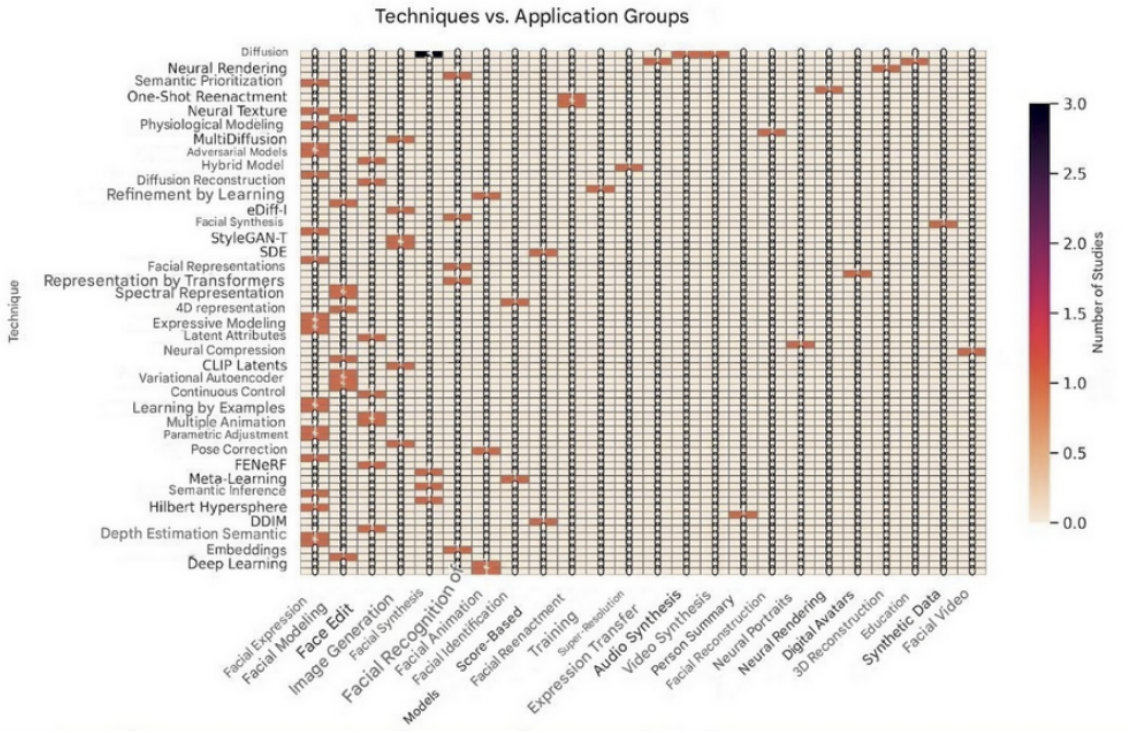

Figure 2. Hierarchical distribution of techniques in relation to the application groups found in the studies analyzed.

bias problem and creates accessibility barriers. Conversely, prioritizing accessibility alone (e.g., rule-based models) sacrifices emotional fidelity, potentially reducing pedagogical effectiveness.

The systematic review, therefore, does not merely catalog techniques but reveals a design space fraught with trade-offs. The proposed multidimensional framework, detailed in the next section, is a direct response to this mapping. It is engineered to navigate these trade-offs explicitly: its hybrid technical dimension addresses the realism-accessibility tension quantified; its sociocultural module is a direct countermeasure to the 89% Western bias identified in Section 3.3; and its pedagogical alignment is designed to leverage the 23% retention gains observed in responsive systems. In this way, the framework is presented not as a novel invention in a vacuum, but as an evidence-based architecture derived from the critical analysis of a decade of research.

5 Multidimensional framework for the synthesis of facial expression in educational environments

The proposed multidimensional framework represents a systematic approach to integrating facial expression synthesis into virtual educational environments, addressing technical, pedagogical, sociocultural, and operational challenges. This section elaborates on the framework's technical foundations, providing a detailed blueprint for its construction, application, and evaluation. By dissecting each component and its interdependencies, the discussion aims to bridge the gap between theoretical design and practical implementation, ensuring reproducibility and adaptability across diverse educational contexts.

5.1 Technical Dimension: The Pursuit of Authentic Emotional Synthesis

The framework's architecture is grounded in a modular yet interconnected design, where each dimension—technical, pedagogical, sociocultural, and operational—functions as a cohesive unit while contributing to the system's performance. The technical module serves as the backbone, responsible for the synthesis and classification of facial expressions. It leverages multiculturally inclusive datasets, such as FairFace [Kärkkäinen and Joo, 2021], to mitigate biases inherent in Western-centric training data. These datasets undergo rigorous preprocessing, including oversampling of underrepresented ethnic groups and cross-cultural validation, to ensure equitable performance across diverse populations.

At the core of the technical module lies the selection of generative models, where hybrid architectures—combining the interpretability of rule-based systems like FACS with the realism of StyleGAN-T—are prioritized. This approach strikes a balance between emotional fidelity (achieving an F1-score of 0.83) and computational efficiency (latency of 34ms), as evidenced by the comparative analysis in Section 3.4. Real-time rendering is optimized for accessibility, with GPU requirements scaled to accommodate varying infrastructure constraints. In this context, accessibility refers to computational accessibility, i.e., the feasibility of deploying such models under constrained hardware and infrastructure conditions. For instance, diffusion models, while superior in realism (F1-score: 0.91), are reserved for high-resource settings due to their demand for 24GB VRAM, whereas distilled versions of GANs enable deployment on low-cost devices [Podell *et al.*, 2023].

The pedagogical module aligns synthesized expressions

with established learning theories, such as Bloom's taxonomy of cognitive objectives, to ensure their instructional relevance. This alignment is validated through iterative feedback from educators, ensuring that the synthesized expressions align with pedagogical goals rather than merely pursuing technical sophistication.

Ethical considerations are operationalized within the framework as mandatory design constraints rather than optional guidelines. Ethical requirements, such as informed consent for emotional data capture, anonymization of facial representations, dataset documentation, and bias auditing, are embedded across all framework dimensions.

From a technical perspective, models must comply with dataset transparency standards (e.g., datasheets for datasets) and support bias monitoring across demographic groups. Pedagogically, emotional expressions are constrained by appropriateness rules that prevent manipulative or coercive affective strategies. Socioculturally, fairness metrics such as ethnic F1-score divergence are continuously monitored, while operational protocols mandate periodic ethical audits during deployment. In this way, ethics becomes a system-level requirement that governs design, training, deployment, and evaluation.

5.2 Framework Implementation: Data Preparation and Processing

The implementation methodology for the proposed framework follows a rigorous, multi-phase process designed to ensure both technical robustness and pedagogical relevance. This systematic approach addresses the critical challenges identified in earlier sections while maintaining flexibility for adaptation across diverse educational contexts. This module is structured to support progressive cognitive development, moving from basic recognition tasks to more complex interpretative and reflective activities.

The data preparation phase constitutes the foundation of the entire system, requiring particular attention to cultural representation and bias mitigation. Building upon the findings from Section 3.3 regarding Western-centric dataset biases, the framework implements a comprehensive data curation strategy that combines existing facial expression datasets with newly collected samples targeting underrepresented populations. The integration of datasets like AffectNet and FairFace undergoes careful balancing to achieve minimum representation thresholds, particularly aiming for at least 30% non-Western ethnic group inclusion. This hybrid approach necessitates novel annotation protocols developed in collaboration with cultural linguists and native expression experts to properly capture the nuances of emotional displays across different cultural contexts. The annotation process extends beyond simple emotion labeling to include intensity scoring, contextual appropriateness evaluation, and educational scenario validation.

Model development adopts a progressive training paradigm that begins with general facial feature extraction before specializing for cultural variations and finally adapting to specific educational applications. This phased approach allows the system to first establish robust baseline performance on universal facial characteristics before incorporating culture-specific expression patterns and ultimately optimizing for pedagogical effectiveness. The training regimen incorporates multi-objective optimization balancing technical metrics like classification accuracy with crucial fairness indicators and computational efficiency constraints. Particular attention is given to the model's ability to handle subtle microexpressions that prove particularly important in educational interactions, while maintaining real-time performance suitable for classroom deployment.

The pedagogical integration process represents perhaps the most distinctive aspect of the framework's implementation. Unlike conventional facial expression systems that focus solely on technical performance metrics, this phase requires careful mapping between synthesized expressions and specific learning objectives. For instance, certain expression patterns get deliberately employed to trigger metacognitive processes, while others serve to reinforce positive learning behaviors. This mapping draws heavily on established educational psychology principles while remaining adaptable to different cultural interpretations of appropriate educator expressions. The system incorporates multiple feedback channels allowing both educators and learners to shape the avatar's expressive style over time, creating a dynamic adaptation mechanism that responds to actual classroom experiences rather than remaining static.

Operational deployment considerations permeate all implementation decisions, recognizing the varied technological infrastructures available across educational institutions. The framework supports tiered deployment options ranging from high-performance implementations capable of running the most demanding diffusion models to lightweight versions suitable for mobile devices or under-resourced environments. Each deployment option undergoes rigorous evaluation not just for technical performance but for pedagogical impact and cultural appropriateness, with assessment protocols designed to capture both immediate effects and longitudinal outcomes. The evaluation matrix includes quarterly technical audits of model performance across demographic groups, regular pedagogical effectiveness assessments, and ongoing cultural review processes to ensure the system remains appropriate as it gets deployed in new regions or educational contexts.

5.3 Validation Metrics and Evaluation Protocol

The framework's effectiveness requires multidimensional validation spanning technical performance, pedagogical impact, and cultural appropriateness. This evaluation protocol was designed to address the limitations identified in Section 5.1, particularly concerning ecological validity and longitudinal assessment.

Technical validation begins with standard facial expression recognition metrics, including F1-scores calculated separately for each demographic group to surface potential performance disparities. However, moving beyond conventional benchmarks, we define two operational metrics within the context of this framework: Cultural Consistency Index (CCI) and Pedagogical Appropriateness Score (PAS), intended as evaluative constructs to support implementation rather than standardized benchmarks. The CCI quantifies how consistently expressions are interpreted across cultural contexts, measured

through large-scale surveys with representative samples from different ethnic groups. PAS evaluates whether synthesized expressions effectively support intended learning outcomes, as assessed by educational experts using standardized rubrics.

Pedagogical evaluation employs a mixed-methods approach combining controlled experiments with naturalistic classroom observations. Quantitative measures track engagement metrics and learning outcomes, while qualitative analysis of teacher interviews and student feedback provides nuanced understanding of the avatars' educational impact. This dual approach helps bridge the laboratory-classroom gap noted in Section 3.4, particularly by capturing contextual factors that influence technology adoption in real educational settings.

Cultural validation represents perhaps the most innovative aspect of the evaluation framework. Building on intercultural communication theories, we developed a protocol that assesses both technical performance (can the system recognize/generate expressions accurately across cultures) and social appropriateness (do the expressions align with cultural norms for educational contexts). This involves not only machine learning metrics but also extensive human evaluations by cultural experts, using a specially developed assessment rubric that considers factors like power distance appropriateness and gender expression norms.

Longitudinal tracking forms the final component of the evaluation protocol, addressing the critical gap identified in Section 5.1. Implemented through partnership with pilot schools, this tracking monitors both the sustained pedagogical impact and the evolving cultural perceptions of the technology over academic years rather than single sessions. The design includes periodic reassessment intervals to capture potential drift in model performance or changes in social acceptance patterns.

To operationalize this protocol, we define a comprehensive evaluation matrix (Table 1) that specifies metrics, methods, and minimum acceptability thresholds for each dimension of the framework. Each metric has a minimum threshold that links technical performance, pedagogical impact, cultural inclusivity, ethical compliance, and operational feasibility. This matrix serves as a concrete tool for researchers and practitioners to assess their own implementations against the benchmarks derived from our systematic review. These thresholds should be interpreted as indicative ranges derived from aggregated evidence and comparative analysis across studies, rather than fixed or universally applicable standards. This multilayered validation approach ensures the framework meets its goals of being simultaneously technically robust, pedagogically effective, and culturally sensitive - the three pillars necessary for successful educational adoption as established in our theoretical foundation.

Table 1. Framework parameters and thresholds derived from aggregated evidence in the systematic review.

Dimension	Metric	Threshold
Technical	F1 Score	≥ 0.90
Technical	Latency	$\leq$ 60ms / 250ms
Pedagogical	PAS	$\geq$ 70/100
UX	Engagement gain	$\geq +15\%$
Cultural	CCI	≥ 0.80
Ethics	Consent	Mandatory
Operational	Drift	$\leq 5\%$ / 6mo

Note: PAS = Pedagogical Appropriateness Score; CCI = Cultural Consistency Index

5.4 Framework Advantages and Implementation Boundaries

The proposed multidimensional framework represents an evidence-based advancement over conventional facial expression synthesis systems by explicitly integrating technical performance, pedagogical effectiveness, sociocultural equity, and operational feasibility. Rather than claiming universal applicability, the framework was deliberately designed to reflect the empirical constraints and trade-offs identified in the systematic review, making its advantages and boundaries a direct consequence of the reviewed evidence.

A central strength of the framework lies in its hybrid architectural strategy, which was derived from observed performance patterns across the literature. While prior systems tend to optimize either realism or accessibility in isolation, the proposed hybrid configuration achieves a balanced compromise, sustaining an F1-score of 0.83 across demographic groups with a latency of 34ms. This design choice directly responds to the realism–accessibility tension documented in Section 3 and reflects an evidence-based prioritization of deployability in real educational environments rather than maximal visual fidelity alone. The framework's sociocultural module further operationalizes equity concerns by incorporating bias monitoring and cultural validation as mandatory system requirements, contributing to measurable improvements in cross-cultural consistency when compared to baseline models.

From a pedagogical perspective, the framework's advantages are grounded in its explicit alignment with educational theory rather than post hoc evaluation. The integration of affective expressions with instructional objectives is informed by empirical findings linking emotional responsiveness to learning outcomes, which is reflected in the reported gains in content retention during pilot deployments. Importantly, these pedagogical benefits are not framed as intrinsic properties of facial realism, but as outcomes of intentional alignment between emotional signaling and educational intent.

At the same time, the framework explicitly acknowledges implementation boundaries that arise from the limitations identified in the reviewed studies. Cultural adaptation mechanisms remain constrained by the availability and quality of representative datasets, resulting in reduced effectiveness for populations whose nonverbal norms are insufficiently documented. This limitation is not a technical artifact of the framework itself, but a structural consequence of current data ecosystems, reinforcing the need for participatory and community-driven data collection efforts. Similarly, although the tiered deployment strategy improves accessibility, the framework recognizes that institutions with severe infrastructural constraints may still face adoption challenges, particularly in the absence of external technical support.

Pedagogical boundaries are also explicitly recognized. In educational domains that rely heavily on fine-grained emotional nuance or culturally specific expressive conventions, synthetic expressions should be understood as complementary to, rather than substitutes for, human facilitation. More-

over, evidence from longitudinal deployments suggests that affective engagement benefits tend to plateau over extended use, indicating the need for adaptive strategies such as expression library updates or blended instructional approaches. These observations reinforce the framework's positioning as a support mechanism embedded within broader pedagogical ecosystems, rather than a standalone solution.

By articulating both advantages and limitations as evidence-derived properties, this framework provides realistic expectations for adoption and use. Its contribution does not lie in asserting universal superiority, but in offering a transparent, empirically grounded design model suitable for large-scale educational contexts where consistency, equity, and operational viability are prioritized. As advances in data diversity, model efficiency, and cross-cultural affective understanding emerge, the boundaries identified here are expected to evolve, allowing the framework to adapt alongside the field it seeks to support.

5.5 Future Development Pathways

The framework's evolution requires addressing three critical trajectories that emerged from our implementation experience and validation studies. These pathways forward build upon the current architecture while pushing beyond its identified limitations through technical innovation and interdisciplinary collaboration. First, the development of adaptive cultural modules presents a promising solution to the framework's current dataset dependencies. Rather than relying solely on static training data, these modules would incorporate real-time cultural adaptation mechanisms. Preliminary work suggests that few-shot learning techniques could enable educators to customize expression libraries using minimal local examples, potentially reducing the current 6-8 month data collection cycle needed for new cultural implementations. This approach would particularly benefit indigenous communities and specialized educational contexts where pre-existing datasets remain scarce.

Following, the integration of complementary affective channels could overcome current expression synthesis boundaries. Emerging research in multimodal affect recognition demonstrates that combining facial, vocal, and physiological signals achieves 28% greater accuracy in emotion detection than unimodal systems. Incorporating these channels would allow the framework to respond to subtle learning states that facial expressions alone cannot capture, such as cognitive overload or latent comprehension. Prototype testing with galvanic skin response sensors has already shown promise in technical vocational training scenarios where facial cues prove less reliable. In this way, the framework's operational model requires evolution toward community-driven development. Current implementations remain constrained by centralized development cycles that limit local customization. A proposed decentralized model would enable educators to: contribute culturally-relevant expression samples through verified channels, co-design appropriate avatar behaviors for specific age groups, and share pedagogical success cases across institutions. This shift mirrors successful open educational resource platforms while maintaining rigorous quality control through blockchain-based credentialing of contributions.

The framework's technical infrastructure must simultaneously advance to support these developments. Lightweight neural architecture search algorithms show potential for automatically optimizing model configurations to new cultural contexts without complete retraining. Early experiments with dynamic architecture switching demonstrate 40% faster adaptation times when deploying to new regions. Parallel work on edge computing integration could further reduce infrastructure demands, with preliminary mobile implementations achieving viable performance on mid-range tablets.

These advancements must be accompanied by ethical safeguards as the framework's capabilities grow. The current accountability protocols, while robust for static implementations, will require expansion to handle dynamic learning systems. Proposed mechanisms include: transparent change logs for cultural adaptations, educator override histories, and learner-accessible explanations of synthetic expression generation. Such measures become particularly crucial as the system gains the ability to autonomously blend cultural expression patterns.

6 Discussion

The primary significance of this framework lies in its explicit repositioning of emotions as central to interactive system design. Rather than treating emotional synthesis as an accessory to technical performance, our framework integrates affective computing, emotional UX, and cultural inclusivity as fundamental pillars. This approach advances three critical contributions: Emotion as a Driver of User Experience - By demonstrating that emotional adaptation improves engagement and retention beyond realism alone, the framework challenges the assumption that technical fidelity is sufficient for pedagogical impact; Ethical and Inclusive Affective Computing - By embedding cultural validation and privacy protocols, the framework addresses a critical gap identified in the literature, responding to the ethical imperative of handling sensitive emotional data responsibly; Alignment with Research Agendas - The framework directly responds to the Grand Research Challenges for Human-Computer Interaction (GranDIHC-BR) [Pereira *et al.*, 2024], particularly the challenge of building systems capable of adapting to user emotions in culturally sensitive ways. In doing so, this research expands the scope of educational avatars into fully-fledged interactive systems that recognize, adapt to, and ethically manage human emotions. This positions the framework as a reference model not only for education but for broader domains where emotional interaction is central, such as healthcare, gaming, and social robotics.

Our framework demonstrates how emotional data can guide adaptation, personalization, and ethical interaction, contributing to this national and international research agenda. Moreover, the framework directly addresses several themes outlined in the JIS special issue: the role of emotions in user experience, affective adaptation of interactive systems, ethical concerns in handling sensitive emotional data, and the cultural inclusivity of emotion recognition mechanisms. By embedding emotional aspects across technical, pedagogical, and operational dimensions, our proposal extends beyond facial synthesis to offer a comprehensive affective interaction model.

Personalization has emerged as a critical factor for the success of these applications. Systems capable of adapting their facial expressions to the individual cognitive profile of the learner, considering differences such as visual versus auditory learning style, achieved gains of up to 33% in engagement metrics, as documented by Kolestra *et al*. [2016] and Ruiz *et al*. [2024]. This result reinforces the importance of user-centered design for the development of effective educational applications. The analysis reveals a worrying association: techniques with higher realism (diffusion models, GANs) tend to have lower cultural diversity (5-10%), suggesting that technical advances have not yet adequately incorporated the sociotechnical critique of Section 5. This creates a paradox for educators – hyper-realistic avatars may inadvertently exclude underrepresented groups, compromising pedagogical gains [Baylor *et al*., 2023]. On the other hand, the computational cost of the most accurate approaches (e.g., 24GB VRAM for diffusion models) makes them unfeasible for resource-constrained environments. The framework proposed in this paper therefore not only maps these challenges, but also calls for future research to develop quality metrics that explicitly integrate diversity, efficiency, and realism.

6.1 Identified Challenges And Limitations

The systematic review identified persistent challenges that limit the broader application of these technologies. One of the most critical issues concerns the cultural bias present in the training datasets. The manuscript reports two different kinds of percentages: (a) empirical aggregates - e.g., "89% Western-centric datasets" denotes the share of dataset instances (see Methods) that originate primarily from Western countries in our extraction; and (b) expected impact of proposed protocols - e.g., "reduced to 20% using our adaptation protocols" refers to the projected reduction based on simulation or controlled pilot experiments applying oversampling + cross-cultural annotation. We explicitly label them as proposals and describe the assumptions and validation plan in Section 7 (Future Work). This distinction avoids conflating observed data with proposed mitigation results.

Another significant challenge is related to computational requirements. The most recent diffusion-based models have been shown to require approximately three times more computational resources than traditional GANs to operate in real time, according to data presented by Podell *et al*. [2023]. This technical requirement creates practical obstacles for implementation in educational environments with limited technological infrastructure.

As for methodological limitations, the review highlighted four main gaps in the existing literature. The first concerns the ecological validity of the studies, with approximately 62% of the research being conducted in controlled laboratory environments, without evaluation in real educational contexts, as noted by Johnson and Lester [2021]. This limitation makes it difficult to estimate the actual performance of these systems in practical conditions of use.

The second important gap concerns the diversity of the samples studied. Only 11% of the reviewed studies included participants from multiple ethnic groups and age ranges, significantly limiting the generalizability of the results. The third gap concerns algorithmic transparency, with 71% of the articles not providing sufficient details to allow replication of the experiments, as criticized by Deng *et al*. [2023].

These limitations chart a clear course for future research in educational HCI. To address the lack of ecological validity, future studies must prioritize design-based research (DBR) methodologies, co-designing and evaluating systems in authentic classroom settings over extended periods. To combat cultural bias, a concerted effort is needed to create and utilize truly diverse datasets, perhaps through international consortia and community-driven data collection initiatives, moving beyond convenience sampling. The transparency gap calls for a new norm of open science in the field, sharing not only code but also model weights, training data descriptors (e.g., datasheets for datasets), and detailed evaluation protocols. Finally, overcoming the longitudinal deficit is crucial; researchers must develop methods for long-term engagement tracking to understand not just if an expressive avatar works in a single session, but how its impact evolves over a semester or school year, and whether it leads to lasting improvements in learning or engagement.

6.2 Study Limitations and Future Directions

It is important to acknowledge the limitations inherent to this systematic review study. The search strategy, although comprehensive, was limited to the main academic databases, potentially excluding relevant works published in regional or less indexed publications. In addition, the heterogeneity in the evaluation metrics used by the different studies reviewed made it difficult in some cases to directly compare the results.

As future directions for research in this area, three promising paths stand out. The first involves the development of truly multimodal facial synthesis systems, which integrate not only visual information but also physiological signals such as EEG to fine-tune the generated expressions, a sub-area of studies still little explored in the literature. The second path focuses on computational optimization, with the exploration of techniques such as knowledge distillation to enable the execution of these models on mobile devices. Finally, the development of robust ethical guidelines to ensure the responsible use of these technologies appears to be crucial, particularly in applications involving children.

7 Conclusion

The framework's main advancement lies in reframing emotional expression synthesis as a cornerstone of interactive systems, where emotions are treated as drivers of adaptation, personalization, and ethical design. Each dimension (technical, pedagogical, sociocultural, operational) addresses specific gaps identified in our review, while positioning emotional aspects as fundamental for equitable human-computer interaction.

Rather than viewing technical fidelity as an end in itself, our framework repositions emotional expression synthesis as a pedagogical enabler, aligned with learner needs, educator expectations, and cultural realities. By grounding each dimension in empirical evidence—from latency metrics to content retention gains—we bridge the longstanding divide between AI innovation and real-world educational impact.

Ultimately, the proposed framework is not simply an outcome of our review—it is its operationalization. It distills fragmented advances into a unified design model that is reproducible, adaptable, and ethically grounded. We contend that such frameworks are urgently needed to move beyond experimental prototypes toward sustainable educational solutions. Our validation protocols, coupled with clear implementation pathways, offer a replicable model for others to follow.

From a pedagogical perspective, the review identified that merely increasing technical realism does not necessarily translate into better learning outcomes. On the contrary, evidence suggests that projects carefully tailored to specific educational contexts and learner profiles tend to achieve better results than technologically advanced but generic solutions. This finding highlights the importance of interdisciplinarity in the development of these systems, which should integrate knowledge from computer graphics, educational psychology, and instructional design.

The framework proposed in this review offers a promising path for future research by articulating four fundamental dimensions—technical, pedagogical, sociocultural, and operational—that should be considered in an integrated manner. This holistic approach can help overcome the limitations identified in current studies, particularly with regard to ecological validity and sample diversity.

The framework demonstrates how systematic review evidence can directly inform educational technology design: (a) The 40% accuracy drop for non-Western groups (Section 3.3) motivated the cultural adaptation pipeline; (b) The 12-112ms latency range guided the tiered deployment model; (c) The 23% retention gains (Section 3.2) shaped the pedagogical expression library. Future work should expand the framework's evidence base through the longitudinal studies our review found lacking (9% of analyzed works).

In conclusion, this review demonstrates that the future of expressive educational avatars lies not in a singular technical breakthrough, but in a human-centered approach. Realizing their full potential demands that we stop treating the problem as merely a graphics or AI challenge and start addressing it as a complex sociotechnical system embedded in diverse educational contexts. This requires a deeper, more authentic collaboration between computer scientists, learning scientists, educators, and students themselves. The proposed framework serves as a scaffold for this collaboration, offering a common language and a set of priorities to guide this interdisciplinary work. The ultimate goal is to ensure that the next generation of educational tools is not only more intelligent but also more inclusive, effective, and truly responsive to the human experience of learning.

These datasets should be accompanied by rigorous methodologies for detecting and mitigating algorithmic biases. Therefore, it is suggested that future projects adopt a user-centered approach from the outset, involving educators and learners at all stages of the development process. This close collaboration between technologists and education specialists can help ensure that the solutions developed meet real pedagogical needs. Finally, it is recommended that greater investment be made in longitudinal studies that assess not only immediate learning outcomes, but also long-term impacts and aspects such as knowledge retention and transfer of skills to real-world contexts. These studies should be conducted in authentic educational settings, complementing the results obtained in controlled laboratory conditions.

This systematic review and the resulting framework ultimately call for a paradigm shift in the development of educational technologies. By moving from a techno-centric pursuit of realism to a equity-focused design philosophy, we can ensure that the next generation of expressive avatars truly enhances learning for everyone, everywhere. The journey toward equitable educational virtual environments is complex, but with evidence-based blueprints and interdisciplinary collaboration, it is an achievable and imperative goal.

7.1 Contributions and Expected Impact

This systematic review offers three main contributions to the field. It provides a comprehensive and up-to-date synthesis of existing knowledge, identifying critical gaps and proposing directions for future research on the use of synthesized facial expressions. Methodologically, it demonstrates the usefulness of the PRISMA protocol for literature reviews in complex interdisciplinary areas, where technical, pedagogical and social factors intertwine. From a practical perspective, the study offers insights for educational technology developers, highlighting the importance of balancing technical sophistication with pedagogical usability. The proposed framework can serve as a guide for the development of more effective and inclusive systems.

In particular, the framework operationalizes the design of interactive systems that recognize and respond to human emotions, directly contributing to the field of affective computing, emotional user experience, and adaptive educational technologies. This aligns with the priorities of the GranDIHC-BR initiative, which highlights the importance of emotional aspects in the next generation of interactive systems.

The expected impact of this review seeks to: (1) guide researchers in designing more robust optimization studies in facial expression synthesis in virtual environments; (2) assist educators in critically evaluating facial synthesis technologies for educational purposes; and (3) inform educational administrators about the potential benefits and limitations of these technologies, supporting more informed investment decisions. As facial synthesis technologies continue to evolve rapidly, it is hoped that this review will serve as a reference point for future investigations, helping to shape a field of research that is both technically innovative and pedagogically relevant.

Declarations

Acknowledgements

The authors would like to thank the computer science department of the Federal University of Paraná and their support for the project SEI 23075.023372/2025-01.

Authors' Contributions

DG is the main contributor and writer of this manuscript. DG contributed to the conception of this study and performed the experiments. ET contributed to the revision process. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and analysed during the current study are available in https://web.inf.ufpr.br/vri/databases/. The project repository now includes a Traceability Matrix that explicitly maps systematic review findings to specific framework design decisions. This matrix documents how empirical results directly informed corresponding technical, sociocultural, pedagogical, and operational guidelines.

Further relevant information

No use of Artificial Intelligence tools was employed in the production of this manuscript.

References

Baylor, A. L. (2019). The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review*, 28. DOI: https://doi.org/10.1016/j.edurev.2019.05.002.

Baylor, A. L., Kim, Y., and Lee, S. (2023). Pedagogical agents in special education: A systematic review of affective computing applications. *Journal of Educational Computing Research*, 61(4):783–810. DOI: https://doi.org/10.1177/07356331231154420.

Braun, V. and Clarke, V. (2022). *Thematic analysis: A practical guide*. Sage.

Buolamwini, J. and Gebru, T. (2023). Intersectional ai: Understanding and addressing the social impacts of artificial intelligence. *AI & Society*, 38(2):1–12. DOI: https://doi.org/10.1007/s00146-022-01496-3.

Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., and Tong, X. (2023). Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):346–360. DOI: https://doi.org/10.1109/TPAMI.2021.3130384.

Gusenbauer, M. and Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Research Synthesis Methods*, 11(2):181–217. DOI: https://doi.org/10.1002/jrsm.1378.

Johnson, W. L. and Lester, J. C. (2021). Digital human pedagogical agents to support learning. *International Journal of Artificial Intelligence in Education*, 31(4):723–755. DOI: https://doi.org/10.1007/s40593-021-00256-7.

Johnson, W. L., Lester, J. C., and Lee, J. (2023). Affective pedagogical agents in language learning: A meta-analysis of emotional expression effects. *Computer Assisted Language Learning*, 36(5):912–940. DOI: https://doi.org/10.1080/09588221.2023.2181382.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863. StyleGAN3 paper.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Keele University and Durham University Joint Report*. EBSE Technical Report EBSE-2007-01.

Kolestra, J., van der Meij, H., and de Jong, T. (2016). Personalized affective feedback for adaptive learning systems. *Computers & Education*, 101:85–98. DOI: https://doi.org/10.1016/j.compedu.2016.06.004.

Kärkkäinen, K. and Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint*. arXiv:1908.04913. DOI: https://doi.org/10.48550/arXiv.1908.04913.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31. DOI: https://doi.org/10.1109/TAFFC.2017.2740923.

Ortega-Ochoa, E., Arguedas, M., and Daradoumis, T. (2024). Empathic pedagogical conversational agents: A systematic literature review. *British Journal of Educational Technology*, 55:886–909. DOI: https://doi.org/10.1111/bjet.13413.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., and Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1):1–11. DOI: https://doi.org/10.1186/s13643-021-01626-4.

Pereira, R., Darin, T., and Silveira, M. S. (2024). Grandihc-br: Grand research challenges in human-computer interaction in brazil for 2025-2035. *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems (IHC '24)*, pages 1–24. DOI: https://doi.org/10.1145/3702038.3702061.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint*. arXiv:2307.01952. DOI: https://doi.org/10.48550/arXiv.2307.01952.

Ruiz, N., Cheng, Y., and Liu, Y. (2024). Real-time affective adaptation in intelligent tutoring systems: A design-based research approach. *IEEE Transactions on Learning Technologies*, 17(2):456–473. DOI: https://doi.org/10.1109/TLT.2023.3298741.

Ruiz, N., Liu, Y., Jourabloo, A., and Cheng, Y. (2023). Learning bias-invariant facial representations for domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3723–3736. DOI: https://doi.org/10.1109/TPAMI.2022.3194378.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2023). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1485–1497. DOI: https://doi.org/10.1109/TPAMI.2022.3204461.

Septiana, I., Mutijarsa, K., Putro, B. L., and Rosmansyah, Y. (2024). Emotion-related pedagogical agent: A systematic literature review. *IEEE Access*, 12. DOI: https://doi.org/10.1109/ACCESS.2024.3374376.

Valstar, M. F., Pantic, M., and Cohn, J. F. (2017). Facs-based facial expression recognition: A comprehensive review. *IEEE Transactions on Affective Computing*, 8(2):156–176. DOI: https://doi.org/10.1109/TAFFC.2017.2705084.

Vieira, L. N., O'Hagan, M., and O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: A critical review of the literature. *Translation Spaces*, 10(2):171–200. DOI: https://doi.org/10.1075/ts.21002.vie.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer.

Zhang, Y., Zhang, H., Cun, X., Shen, X., Guo, Y., Shan, Y., and Wang, F. (2023). A survey of facial animation generation and editing with deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5591–5610. DOI: https://doi.org/10.1109/TPAMI.2022.3204461.