Automatic Creation of an AI Presenter using an Educator's Photo

Naman Chawla[1], Anuradha Misra[2][0000-0002-3790-8798]
Amity School of Engineering & Technology, Amity University Uttar Pradesh, India

Abstract: Portrait animation has undergone a transformative evolution, transitioning from traditional diffusion-based frameworks to efficient methods that leverage implicit keypoint representations. These advancements have enabled realistic facial animations with greater computational efficiency and controllability, addressing challenges in achieving real-time performance. LivePortrait represents a cutting-edge framework that combines implicit keypoint optimization, stitching modules, and landmark-guided retargeting to create lifelike animations. This report synthesizes the latest research and comparative studies on portrait animation technologies, with a detailed focus on LivePortrait's methodologies and practical implementations. Furthermore, this report explores enhancements to bridge the gap between static imagery and realistic avatars, paving the way for advanced applications in entertainment, education, and virtual communication. By adhering to a Scopus-style format, this document aims to provide a comprehensive resource for researchers and developers.

1 INTRODUCTION

The field of portrait animation has made significant strides in recent years, with innovations that transform static images into lifelike, dynamic avatars. The central objective of this technology is to recreate natural facial expressions and movements while ensuring high-quality output and real-time processing capabilities. This evolution is particularly evident in the transition from diffusion-based frameworks, which rely on iterative refinement, to models that utilize implicit keypoints for efficient motion transfer.

LivePortrait emerges as a robust framework that addresses key challenges in this domain, such as computational efficiency, controllability, and realism. By optimizing implicit keypoints and incorporating advanced modules like stitching and retargeting, LivePortrait achieves seamless animations with minimal artifacts. This approach not only improves generalization across diverse facial structures but also allows fine-grained control over expressions and motion dynamics.

The importance of portrait animation extends beyond academic research, finding applications in entertainment (e.g., gaming and film production), education (e.g., interactive learning platforms), and telecommunication (e.g., virtual meetings and avatars). As this field continues to evolve, experimenting with new methods and enhancing existing frameworks remains critical to meeting the ultimate goal: bringing static images to life as realistic avatars capable of complex, human-like interactions.

This report delves into the methodologies, comparative analyses, and practical implementations of LivePortrait and related technologies. It aims to highlight the innovations that make LivePortrait a benchmark in portrait animation while identifying areas for future research and development.

2 LITERATURE REVIEW

The advent of artificial intelligence (AI) has significantly transformed various sectors, with education being a particularly prominent area of innovation. One of the most compelling advancements in this domain is the development of AI-driven portrait animation technologies, which enable the creation of lifelike avatars from static images. These avatars, commonly referred to as "talking heads," are digital constructs capable of mimicking human speech and facial expressions in real time. They present an exciting opportunity to revolutionize how educational content is delivered by allowing educators to develop personalized, expressive, and engaging AI avatars that can dynamically convey information to learners. This literature review thoroughly explores the evolution of portrait animation, with a specific focus on LivePortrait and a range of influential open-source libraries. It also examines their potential and practical implications for transforming the educational landscape.

2.1 Introduction to Portrait Animation

Portrait animation is a subfield of computer vision and deep learning that centers on converting static images into

animated representations by simulating facial movements and expressions. These animations are often synchronized with audio inputs to generate highly realistic talking face sequences. The applications of portrait animation are diverse, encompassing sectors such as entertainment, digital communication, virtual and augmented reality, and most notably, education. The key objective is to synthesize facial behaviors that are temporally coherent, identity-preserving, and semantically aligned with spoken content, thereby creating engaging and human-like digital interactions.

2.2 Evolution of Portrait Animation Techniques

Over the past decade, portrait animation techniques have undergone substantial evolution. The earliest models relied heavily on graphical heuristics and manual animation, while contemporary methods now leverage powerful deep learning architectures and large-scale datasets. This section outlines the development of the main categories of techniques used in portrait animation.

2.2.1 Diffusion-Based Methods

Diffusion-based models represent a cutting-edge paradigm in generative modeling. They work by progressively denoising a random signal to generate detailed and high-fidelity outputs. In the context of facial animation, these models offer superior visual quality and a high degree of expressiveness, though they often come with substantial computational requirements.

• FADM (Facial Animation Diffusion Model): This method employs a combination of implicit keypoints and 3D morphable models (3DMMs) to produce highly realistic facial animations. The integration of 3DMMs helps preserve the subject's identity and expression continuity. However, due to the multiple sampling steps involved, FADM is computationally intensive and may not be suitable for real-time deployment without significant optimization.

• Face Adapter: Face Adapter enhances traditional diffusion models by introducing identity-preserving adapters and spatial condition generators. These modules ensure that the output animation maintains consistent facial characteristics and adapts well to various head poses and expressions. Despite its effectiveness in preserving visual identity, its high inference cost limits its scalability for widespread real-time applications.

• AniPortrait: AniPortrait achieves greater animation control by integrating explicit spatial conditions, including facial keypoints, masks, and expression labels. These enhancements allow for more accurate and expressive facial reenactment. However, like other diffusion-based approaches, AniPortrait demands considerable computational power and access to expansive annotated training data.

2.2.2 Implicit Keypoint-Based Methods

To overcome the limitations of diffusion-based models, researchers have developed implicit keypoint-based methods. These models are designed to identify latent keypoints from a static image or short video and use them to control facial deformation and motion synthesis in a computationally efficient manner.

- First-Order Motion Model (FOMM): FOMM is a pioneering framework that introduces the concept of local affine transformations around dynamically learned keypoints. Using a first-order Taylor expansion, it models motion between the source and driving frames. While it is significantly faster than diffusion models and suitable for real-time use, FOMM struggles with modeling fine-grained facial nuances, particularly during exaggerated expressions or large pose shifts.

- Face Vid2Vid: An extension of FOMM, Face Vid2Vid enhances the motion representation by incorporating 3D implicit keypoints and pose estimation. This method provides the flexibility to animate faces across varying viewpoints. Despite its improvements, Face Vid2Vid still lacks precise expression modeling and can suffer from identity drift in long sequences.

- Thin-Plate Spline Motion Model (TPSM): TPSM offers a non-rigid transformation framework using thin-plate spline warping to simulate complex facial movements. This model is particularly useful in capturing subtle and exaggerated facial gestures like eyebrow raises and cheek puffs. Nevertheless, it is sensitive to parameter configuration and may introduce spatial artifacts if improperly tuned.

2.3 Hybrid Approaches

Hybrid models seek to combine the expressive power of diffusion techniques with the real-time capabilities of keypoint-based models. These frameworks aim to deliver a balanced trade-off between realism, efficiency, and control.

• LivePortrait: LivePortrait exemplifies a successful hybrid strategy. It employs an implicit keypoint framework while introducing novel stitching and retargeting modules that provide superior control over facial expressions. Unlike traditional systems, LivePortrait scales up its training regime to 69 million high-resolution video frames, benefiting from a diverse image-video joint training strategy. Its motion transformation modules are enhanced through optimized loss functions and upgraded network designs. Notably, the system achieves a high generation speed of 12.8 milliseconds per frame on an RTX 4090 GPU, making it practical for real-time educational applications. LivePortrait's stitching and MLP-based retargeting also enable fine-tuned control over features like lip movement, gaze, and head orientation.

2.4 Open-Source Libraries for Avatar Creation

A wide array of open-source projects are contributing significantly to democratizing the development of talking-head avatars. These tools are particularly valuable for educators and developers looking to integrate expressive virtual characters into learning environments without extensive technical expertise.

2.4.1 SadTalker

SadTalker supports the generation of 3D motion coefficients from single static images using audio-driven expression modeling. It uses pre-trained 3D face reconstructions and neural rendering pipelines to create avatars that are both expressive and temporally consistent. Its open-source implementation allows for flexibility in integration and customization.

2.4.2 AudioGPT

AudioGPT is a multimodal AI framework that extends language model capabilities into the audio domain. It supports tasks such as speech-to-text, speech generation, music synthesis, and talking-head generation. The system combines speech recognition with emotional tone and semantic alignment, enabling more meaningful and engaging avatar animations.

2.4.3 LiveTalking

LiveTalking offers real-time interactive digital human avatars with integrated lip-sync and audio processing. Its focus on low-latency performance makes it suitable for live classrooms and webinars. Features include head movement tracking, voice modulation, and real-time feedback, enhancing the sense of presence in virtual teaching.

2.4.4 EchoMimic

EchoMimic employs editable landmark conditioning and landmark-to-frame translation to drive portrait animations from audio signals. It is optimized for generating emotionally expressive avatars and supports advanced editing features such as lip-only control or expression freezing, making it highly adaptable for personalized teaching avatars.

2.4.5 Thin-Plate Spline Motion Model

The TPS motion model is also implemented as a standalone open-source tool for generating facial animations from static images. It offers a lightweight and fast animation pipeline, suitable for applications where computation is constrained but animation fidelity is still desired.

2.5 Impact on the Educational Sector

The educational sector stands to benefit immensely from AI-driven portrait animation systems, especially in the era of remote and hybrid learning. These tools can be leveraged to deliver content more interactively, inclusively, and effectively.

2.5.1 Enhancing Engagement and Interactivity

The integration of animated avatars into educational content can significantly enhance student engagement. Unlike static lectures or slides, animated presenters that display facial cues, lip movements, and expressions help simulate human interaction, making content delivery more relatable. This visual dynamism is especially helpful for younger learners and students in remote learning environments where attention retention is critical.

2.5.2 Personalization of Learning Content

Talking avatars can be personalized to reflect cultural, linguistic, or age-specific characteristics, improving relatability for diverse student populations. Educators can create customized AI avatars that match their own voice, appearance, or teaching style, thus maintaining consistency in instructional delivery while benefiting from automation.

2.5.3 Bridging Language Barriers

Real-time multilingual translation and speech synthesis embedded in AI avatars can break down language barriers. This is especially useful in international schools or multicultural classrooms. By integrating language models

and lip synchronization engines, a single avatar can switch between languages fluidly while maintaining accurate phoneme-viseme alignment.

2.5.4 Consistency and Availability

AI avatars offer the advantage of consistent performance and round-the-clock availability. Unlike human educators, they can deliver repeated content with uniform quality, enabling asynchronous and self-paced learning. Students can interact with these avatars at any time to revise concepts or revisit lessons, making learning more flexible.

2.5.5 Challenges and Considerations

Despite their advantages, the deployment of AI avatars in educational settings must consider ethical, technical, and pedagogical implications. Challenges include ensuring the factual accuracy of AI-generated content, mitigating algorithmic bias in facial animations, and preserving the empathetic dimension of human teaching. It is also crucial to assess how these systems affect student-teacher relationships and to develop safeguards against over-reliance on automation. As noted by education policy experts, while AI has the potential to supplement and even enhance traditional teaching, it must be implemented as a complement to—not a replacement for—human educators.

In summary, AI-driven portrait animation technologies are poised to redefine educational content delivery by making it more engaging, inclusive, and adaptive. With ongoing advancements in both the underlying algorithms and available tools, these systems offer transformative potential—but their integration must be approached thoughtfully to ensure equitable and effective learning outcomes.

3 METHODOLOGY

3.1 LivePortrait Framework

The LivePortrait framework represents a significant leap forward in portrait animation, introducing a suite of innovative techniques that enhance both performance and realism. It is specifically designed to overcome limitations in earlier models that struggled with balancing computational efficiency, animation fidelity, and adaptability to diverse facial expressions and poses. LivePortrait not only addresses these challenges but also lays the groundwork for scalable, real-time avatar animation, making it an optimal choice for applications such as education, digital communication, and virtual assistants.

- Implicit Keypoint Optimization: At the heart of LivePortrait lies its novel implicit keypoint optimization strategy. Rather than relying on manually annotated or static keypoints, the system learns to infer dynamic keypoints that adapt to each frame of input. This approach enables more nuanced motion capture, allowing the model to better reflect the subtleties of human expressions. The keypoints are optimized in a latent space, facilitating continuous updates during inference, which ensures smoother and more responsive animations. Furthermore, this strategy significantly reduces the need for manual preprocessing, accelerating both training and deployment.

- Stitching and Retargeting Modules: LivePortrait introduces specialized modules for stitching and retargeting that work in tandem to improve coherence across frames and personalize animations. The stitching module mitigates temporal inconsistencies that often arise in frame-by-frame generation, especially in sequences involving rapid head or eye movements. It does so by examining the temporal dynamics of facial regions such as the eyes, mouth, and jawline, then applies learned adjustments to ensure continuity. The retargeting module, on the other hand, allows for adaptive modulation of specific facial features. For instance, educators can customize avatars to maintain consistent eye contact or emphasize particular expressions while speaking. This level of control is particularly beneficial in applications demanding emotional intelligence or precise lip synchronization, such as virtual instructors or AI narrators.

- Data Scaling: One of LivePortrait's most powerful assets is its large-scale training dataset, which encompasses 69 million high-resolution facial frames. This dataset spans a wide demographic range and includes a diverse array of facial structures, ethnicities, emotional states, and environmental conditions. To maximize the generalizability of its models, LivePortrait employs sophisticated data augmentation techniques, such as simulated occlusions, lighting variations, and pose shifts. These strategies not only improve robustness but also ensure the system can handle edge cases like partially occluded faces or unconventional expressions. Additionally, the dataset's sheer scale allows the

model to learn rare motion patterns, improving accuracy in both common and atypical animation scenarios.

3.2 Key Features

- Stitching: The stitching module plays a critical role in maintaining visual continuity across generated frames. Unlike simple smoothing filters, it incorporates a temporal attention mechanism that identifies and corrects frame-to-frame discrepancies. This includes compensating for sudden changes in head orientation, blinking, or lip shape that might otherwise result in jittery animations. The module leverages recurrent neural network layers to model frame transitions, ensuring that facial regions remain spatially and temporally consistent. By focusing particularly on traditionally challenging areas like hair contours, ear alignment, and jaw articulation, stitching helps produce videos that are not only realistic but also pleasant to watch over long sequences.

- Lip Retargeting: Lip synchronization is a cornerstone of any believable talking avatar. LivePortrait's lip retargeting engine uses a dual-stage pipeline that first extracts phoneme-aligned lip motions from the driving video and then maps them onto the target face using a learned deformation model. This process is fine-tuned through perceptual loss functions to match the rhythm, intensity, and articulation style of the speaker. The system accounts for variations in lip thickness, curvature, and protrusion, making it capable of animating a wide variety of facial morphologies without introducing visual distortions. Furthermore, lip movements are aligned not just to phonemes but also to prosodic cues like pitch and tempo, enhancing the emotional richness of speech.

- Landmark-Guided Optimization: To ensure precision and realism in facial expression rendering, LivePortrait employs a landmark-guided optimization module. This component identifies key facial landmarks, such as the corners of the mouth, inner and outer eye canthi, and nose bridge, and uses them as spatial anchors for animation. These landmarks guide the adjustment of motion vectors, ensuring that micro-expressions—like a smirk, squint, or brow raise—are accurately reflected in the animated output. The system dynamically weights the influence of each landmark based on contextual cues, allowing for emphasis on different facial zones depending on the emotional tone or focus of the animation. For example, during a serious explanation, the system might highlight eyebrow and eye movements, whereas during a joyful moment, it might enhance cheek and lip activity.

In summary, LivePortrait's methodology combines large-scale data training, intelligent motion optimization, and modular feature control to create a highly adaptable and realistic animation framework. Its techniques support both the high-fidelity requirements of media production and the real-time demands of educational and interactive applications, positioning it as a versatile solution in the evolving landscape of digital avatars.

4 EXPERIMENTAL RESULTS

4.1 Performance Metrics

LivePortrait achieves a remarkable per-frame generation speed of 12.8 ms on an RTX 4090 GPU, enabling real-time animation capabilities that far exceed conventional diffusion-based methods, which often require hundreds of milliseconds to seconds per frame due to iterative denoising steps.

- Controllability: By leveraging a compact implicit-keypoint representation combined with bespoke stitching and retargeting modules, LivePortrait affords pixel-level control over distinct facial regions (eyes, lips, brows). These modules use lightweight multilayer perceptrons (MLPs) to adjust motion vectors without significant overhead, resulting in seamless editing of expression intensity, gaze direction, and lip articulation.
- Efficiency: The underlying implicit-keypoint framework removes the need for expensive 3D morphable model fits or extensive iterative sampling. Compared to diffusion models such as AnimateMe, which employ multi-step diffusion processes often taking over 100 ms per frame, LivePortrait's direct regression pipeline reduces compute by up to 10×, permitting 24 fps or higher on high-end GPUs.
- Realism: Qualitative evaluations and user studies report that LivePortrait maintains high fidelity in identity preservation and micro-expression detail, with minimal artifacts such as jitter or texture blurring. Even under challenging poses (up to 45° head rotation) and varied lighting, the model produces coherent lip sync and natural skin shading, outperforming earlier first-order models and explicit mesh approaches.

Together, these metrics underscore LivePortrait's balanced optimization of speed, fidelity, and user-driven control—critical for applications in virtual teaching, live streaming avatars, and interactive educational content.

4.2 Comparative Analysis

Method	Framework	Intermediate Motion	Controllability	Efficiency
FOMM [11]	Non-Diffusion	Implicit Keypoints	Low	Moderate
AniPortrait [12]	Diffusion	Explicit Keypoints	Moderate	Low
LivePortrait	Non-Diffusion	Implicit Keypoints	High	High

5 APPLICATIONS

5.1 Real-Time Use Cases

- Entertainment: LivePortrait's ultra-low latency (12.8 ms/frame) makes it well-suited for integrating lifelike facial animations into interactive media. In gaming and virtual production, real-time facial motion capture systems render high-fidelity 3D avatars directly from binocular video streams at over 60 fps, eliminating costly offline reconstruction. Advanced audio-driven algorithms dynamically drive character expressions in VR and AR applications, enabling seamless lip-sync, gaze shifts, and micro-expressions that boost immersion.

- Education: AI-driven avatars are revolutionizing e-learning by providing interactive virtual instructors that adapt in real time to pedagogical needs. Generative AI avatars can process spoken queries, generate context-aware responses via large-language models, and speak with synthetic voices, creating a responsive learning environment that surpasses static video lectures. Experimental studies have shown that virtual avatars in educational videos significantly improve user engagement and knowledge retention compared to traditional video content.

- Telecommunication: In virtual meetings, LivePortrait's streaming-capable pipeline supports low-bandwidth avatar transmission without perceptual lag. Live speech-driven telepresence research demonstrates that avatar mediation can convey head nods, facial expressions, and turn-taking cues purely from audio and minimal video inputs, preserving conversational dynamics across distances. Similarly, life-size 2D video avatars in head-mounted display settings maintain co-presence and conversational fluidity, balancing fidelity and immersion.

- Healthcare: Avatar technology is gaining traction in therapeutic contexts, from psychosis intervention to cardiac telemedicine. Therapy trials involving patients dialoguing with personalized avatar embodiments of their auditory hallucinations report significant symptom reduction and higher engagement levels. In telemedicine, nurse-like avatars integrated into remote monitoring platforms improve patient satisfaction and self-care adherence among heart failure patients.

- Marketing and Branding: Brands leverage real-time avatars as virtual spokespeople and digital influencers. Campaigns using volumetric hologram avatars in event kiosks greet attendees and deliver personalized product demos, while marketing research indicates that interactive avatars increase brand recall and emotional connection by substantial margins compared to static advertisements.

- Customer Support: AI avatars offer 24/7 conversational assistance via multimodal interaction. Digital avatar chatbots integrating lip-sync and facial gestures enhance perceived empathy and issue resolution rates, with real-time text-to-speech avatars achieving notable improvements in customer satisfaction over text-only bots.

- Social Media and Content Creation: Influencers and creators use LivePortrait-style tools to produce animated talking-head videos for social platforms. Real-time avatar filters and expression-guided overlays boost viewer retention and engagement metrics by significant percentages.

5.2 Integrations with LivePortrait

- LatentSync by ByteDance Integration: Synchronizing latent space embeddings between audio encoders and facial decoders enhances lip-sync accuracy and emotional nuance, enabling richer audio-driven animations without retraining base networks.

- Full-Body Motion Detection Libraries: Coupling LivePortrait's facial pipeline with OpenPose's real-time multi-person 2D pose estimation (body, hand, and face) facilitates end-to-end full-body avatar animation, operating at over 25 fps on consumer GPUs.

- Real-Time Full-Body from Sparse Data (Mimic Motion): A sparse IMU-and-video framework drives avatars' body kinematics in low-bandwidth settings, combining minimal sensor data with video keypoints for drift-resistant motion capture.

- Expression and Pose Estimation Enhancements: Advanced neural modules using hierarchical motion dictionaries in RGBD space can augment LivePortrait's micro-expression fidelity and support arbitrary avatar topologies.

6 CONCLUSION AND FUTURE DIRECTIONS

LivePortrait represents a significant advancement in portrait animation technology. By addressing the limitations of traditional models, it provides a robust solution that balances quality, efficiency, and controllability. Key achievements include:

- High performance in real-time scenarios.
- Enhanced control over facial expressions and motions.
- Seamless integration of stitching and retargeting modules.

As the field of portrait animation continues to evolve, several areas present promising opportunities for further advancement and innovation. To ensure real-world adaptability, broader applicability, and increased robustness, the following future directions have been identified for systems like LivePortrait and similar audio-visual models:

6.1 Expanding Dataset Diversity: Toward Cross-Cultural Generalization

To build AI systems that are fair, reliable, and globally deployable, it is essential to train on datasets that reflect real-world diversity. Current portrait animation models may exhibit performance discrepancies across ethnicities, age groups, and environmental conditions due to limited training diversity. Enhancing dataset diversity involves incorporating a broad range of facial types, expressions, cultural behaviors, and languages.

By training on datasets containing people from various ethnic backgrounds, age groups, facial geometries, lighting environments, and emotional contexts, models can learn generalized representations of human expressiveness. This results in improved accuracy and fairness across diverse demographics.

Furthermore, including data from different cultural expression norms—such as variations in smiling, eye contact, or head nodding—ensures that the model does not inadvertently reinforce cultural biases. The inclusion of multi-language and multi-accent data also enhances performance in audio-to-lip synchronization across global contexts.

6.2 Advanced Cross-Identity Animation: Enhancing Expression Transfer Between Different Faces

Cross-identity animation enables the application of expressions, lip movements, and head motions from one person (the driving subject) to another individual (the target identity). This has numerous applications in fields such as digital avatars, character puppeteering, virtual teachers, and dubbing for media localization.

However, transferring facial motion between individuals with significantly different facial structures presents a major technical challenge. The model must map expressions accurately while preserving the identity of the target face and avoiding anatomical distortions.

Future improvements in this domain involve:

Using 3D Morphable Models (3DMMs) to normalize differences in facial geometry, employing disentangled latent representations to separate identity from expression, incorporating adaptive expression transfer mechanisms to retain both realism and emotional fidelity.

Such advancements would allow LivePortrait-like systems to generate highly expressive, identity-consistent animations across a broad range of avatars, including stylized or non-human characters.

6.3 Audio-to-Lip Synchronization: Multilingual and Adaptive Lip Movement Modelling

Accurate lip synchronization is a critical requirement for believable portrait animation, especially in applications involving education, communication, and dubbing. While current models perform well in English, extending this performance to multilingual contexts introduces new complexities due to variations in phoneme structures, coarticulation patterns, and lip shapes across languages.

To address these challenges, future models should incorporate:

- Language-aware viseme generation mechanisms,
- Adaptive phoneme-to-viseme mapping for each supported language or dialect,
- Multilingual datasets with aligned audio-video segments to train phonetic and visual correlations effectively.
- Incorporating auxiliary modules from speech recognition systems, such as phoneme alignment models, can improve timing and articulation accuracy. This ensures that animated characters produce lip motions that are not only realistic but also phonetically accurate for different languages and accents.

6.4 Enhanced Robustness: Improving Performance in Challenging Scenarios

In real-world applications, input data is often imperfect due to occlusions (e.g., hands, objects, or masks covering parts of the face), extreme head poses, or poor video quality. Such conditions can disrupt facial landmark detection and degrade the quality of animation outputs.

Improving robustness involves making the model resilient to:

- Occlusions: By training on synthetic and real-world occlusion datasets, the system can learn to infer missing facial information from context.
- Extreme Poses: Pose normalization and view-invariant modeling help maintain consistent outputs even when the subject is not front-facing.
- Noise and Blur: Integrating denoising modules and temporal smoothing techniques ensures stable frame-to-frame coherence.
- Low-resolution Input: Using super-resolution preprocessing, or leveraging depth estimation, improves the model's performance on degraded input data.

Together, these improvements allow portrait animation systems to function reliably across a range of environments, including live video conferencing, mobile applications, and educational content recorded under varying conditions.

REFERENCE

[1] Lu, Y., Chai, J., & Cao, X. (2021). Live speech portraits: real-time photorealistic talking-head animation. ACM Transactions on Graphics (ToG), 40(6), 1-17.

[2] Croitoru, F. A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9), 10850-10869.

[3] Croitoru, F. A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9), 10850-10869.

[4] (N.d.). Rochester.edu. Retrieved June 2, 2025, from https://www.cs.rochester.edu/u/lchen63/siggraph21.pdf

[5] Kyteas, D. (2023, October 23). Understanding real-time facial animation..80.Lv; 80lv. https://80.lv/articles/understanding-real-time-facial-animation/

[6] Fink, M. C., Robinson, S. A., & Ertl, B. (2024). AI-based avatars are changing the way we learn and teach: benefits and challenges. Frontiers in Education, 9. https://doi.org/10.3389/feduc.2024.1416307

[7] Zhang, R., & Wu, Q. (2024). Impact of using virtual avatars in educational videos on user experience. Scientific Reports, 14(1), 6592. https://doi.org/10.1038/s41598-024-56716-9

[8] Jin, A., Deng, Q., & Deng, Z. (2022). A live speech-driven avatar-mediated three-party telepresence system: Design and evaluation. Presence (Cambridge, Mass.), 29, 113–139. https://doi.org/10.1162/pres_a_00358

[9] Wang, X., Zhang, W., Sandor, C., & Fu, H. (2024). Real-and-present: Investigating the use of life-size 2D video avatars in HMD-based AR teleconferencing. In arXiv [cs.HC]. http://arxiv.org/abs/2401.02171

[10] Garety, P. A., Edwards, C. J., Jafari, H., Emsley, R., Huckvale, M., Rus-Calafell, M., Fornells-Ambrojo, M., Gumley, A., Haddock, G., Bucci, S., McLeod, H. J., McDonnell, J., Clancy, M., Fitzsimmons, M., Ball, H., Montague, A., Xanidis, N., Hardy, A., Craig, T. K. J., & Ward, T. (2024). Digital AVATAR therapy for distressing voices in psychosis: the phase 2/3 AVATAR2 trial. Nature Medicine, 30(12), 3658–3668. https://doi.org/10.1038/s41591-024-03252-8