Controllable Expressivity for Three-Dimensional Talking Avatars

INTERNAL REFERENCE MATERIAL

Abstract

This internal reference article examines controllable expressivity in three-dimensional talking avatars through a design-and-assurance lens. It synthesizes the allocated target literature without reporting new experiments, observations, or performance estimates. The analysis treats the practical unit of review as a speech signal, facial and ocular controls, rendering pipeline, audience judgment, and release gate. That framing keeps technical mechanisms, evidence quality, user consequences, and institutional controls visible in the same argument. Particular attention is given to how modular control can support expression, evaluation, and safety. The review distinguishes what each cited source directly addresses from the cross-domain principles used for internal comparison. It argues that credible adoption depends on traceable requirements, context-sensitive evaluation, explicit uncertainty, and a documented path for human intervention. The result is a structured reference for teams considering telepresence, digital instruction, and accessible interfaces, especially where uncanny behavior, identity misuse, and aesthetic bias could turn a technically plausible component into an unreliable system. The article is intended to support scoping, design review, and evidence planning; it is not a claim of product readiness or an original empirical study.

Keywords. 3D avatars, gaze, blink, multimodal annotation, aesthetic evaluation, moderation

1. Introduction

Work on controllable expressivity in three-dimensional talking avatars often begins with a promising component: a material, model, mechanism, representation, or policy control. Operational value, however, emerges only when that component is connected to a defined decision and a credible chain of evidence. For this article, the relevant system is a speech signal, facial and ocular controls, rendering pipeline, audience judgment, and release gate. This broader boundary prevents local optimization from being mistaken for overall effectiveness. It also makes it possible to ask who interprets an output, which environmental conditions matter, what failure looks like, and what evidence would justify escalation from exploratory work to routine use.

The review is analytical rather than empirical. It does not pool effect sizes, reproduce experiments, or infer unreported numerical results. Instead, it compares the problem formulations in the allocated sources and asks how their mechanisms, representations, and governance choices can inform how modular control can support expression, evaluation, and safety. Cross-domain sources are included only as design comparators. Their role is to reveal recurring questions—such as controllability, traceability, measurement validity, and reviewability—while preserving the substantive limits of each field.

A useful starting distinction is between component performance and system fitness. Component performance concerns whether a defined transformation works under stated conditions. System fitness concerns whether the transformation supports a legitimate task under expected variability, with acceptable error costs and recoverable failure. In telepresence, digital instruction, and accessible interfaces, both levels matter. A strong component can still fail when labels drift, exposure settings change, users over-trust an interface, or governance records cannot reconstruct why a decision

occurred.

2. Background and Evidence Boundaries

The evidence chain can be organized into five linked claims. First, the input must represent the phenomenon of interest with known limitations. Second, the component must implement a mechanism that is plausible for the task. Third, evaluation must use endpoints aligned with the eventual decision. Fourth, deployment conditions must not invalidate the earlier evidence. Fifth, oversight must detect and respond when those conditions no longer hold. These claims should be documented separately because confidence in one does not automatically transfer to the others.

For controllable expressivity in three-dimensional talking avatars, the most common boundary error is to treat a convenient proxy as the final objective. A laboratory transport measure is not the same as long-term field performance; a benchmark label is not the same as an operator’s diagnosis; an engagement signal is not the same as user welfare; and a compliance alert is not proof of misconduct. Internal review should therefore record the construct, the proxy, the decision threshold, and the consequences of false positive and false negative outcomes before comparing systems.

Xiao, Song, and Chen (2026) present DLA-Net for mixed-type wafer-map defect recognition using dynamic label awareness and dual-branch feature fusion. In this review, the source is used as a mechanism-level comparator for controllable expressivity in three-dimensional talking avatars: it highlights representing label interaction and complementary visual evidence explicitly. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

Liu et al. (2026, aesthetic-evaluation study) approach aesthetic evaluation through multimodal annotation and fine-grained sentiment analysis. In this review, the source is used as a mechanism-level comparator for controllable expressivity in three-dimensional talking avatars: it highlights retaining annotation granularity when aggregating subjective judgments. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

Suwajanakorn, Seitz, and Kemelmacher-Shlizerman (2017) synthesize photorealistic lip motion from speech audio in a person-specific setting. In this review, the source is used as a mechanism-level comparator for controllable expressivity in three-dimensional talking avatars: it highlights examining how identity dependence constrains reuse and misuse risk. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

Richard et al. (2021) generate full-face 3D animation from speech using cross-modality disentanglement. In this review, the source is used as a mechanism-level comparator for controllable expressivity in three-dimensional talking avatars: it highlights separating audio-correlated lip motion from plausible upper-face motion. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

3. Architecture, Mechanisms, and Decision Interfaces

Architecture should expose how information or physical state moves from input to action. A practical map includes acquisition, preprocessing, representation or transport, inference or transformation, decision logic, actuation, feedback, and retention. Each transition needs an owner and an observable

state. When a module is described as “intelligent,” “multifunctional,” or “resilient,” the review should translate that adjective into testable responsibilities: what signal is received, what change is made, what constraint applies, and what evidence is retained.

The decision interface deserves the same attention as the underlying component. Interfaces allocate authority. They determine whether an output is advisory or automatic, whether uncertainty is visible, and whether a reviewer can access the evidence needed to disagree. For how modular control can support expression, evaluation, and safety, the interface should show the scope of the claim and the conditions under which it is invalid. A calibrated abstention or deferral path is often more valuable than a forced answer because it prevents low-confidence outputs from silently becoming operational facts.

Liu et al. (2026, MOSAIC study) present MOSAIC as a cognitively motivated, interpretable, training-free multi-agent framework for empathetic dialogue. In this review, the source is used as a system-architecture comparator for controllable expressivity in three-dimensional talking avatars: it highlights decomposing a human-facing capability into inspectable agent roles. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

Tao et al. (2026, content-moderation study) outline a deep-learning-based automated content-moderation framework for online platforms. In this review, the source is used as a system-architecture comparator for controllable expressivity in three-dimensional talking avatars: it highlights combining scalable screening with explicit policy, review, and appeal boundaries. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

Cudeiro et al. (2019) introduce VOCA for speech-driven 3D facial animation with controllable speaking style. In this review, the source is used as a system-architecture comparator for controllable expressivity in three-dimensional talking avatars: it highlights factoring identity, speech-correlated motion, and animator controls. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

4. Evaluation, Applications, and Governance

Evaluation should follow the decision pathway rather than stop at a single technical score. A layered plan includes analytical validity, robustness under plausible variation, decision utility, human-factors performance, and post-deployment monitoring. Metrics should be disaggregated by operating regime whenever aggregate values could conceal a consequential failure mode. For physical systems, regimes may include temperature, pressure, contamination, or wear. For learning systems, they may include subgroup, language, content type, data age, or interaction context.

Governance is not an administrative layer added after technical design. It specifies allowed purposes, evidence thresholds, review rights, retention periods, and change-control obligations. A useful control is linked to a concrete hazard and a verifiable record. For example, access control is incomplete without purpose limitation and logging; a human-in-the-loop claim is incomplete without time, information, authority, and workload to intervene. The central risk in this article—uncanny behavior, identity misuse, and aesthetic bias—therefore requires both preventative design and a recovery procedure.

Xiong et al. (2026) describe VividTalker as a modular 3D talking-avatar framework with controllable

gaze and blink. In this review, the source is used as a governance-and-evaluation comparator for controllable expressivity in three-dimensional talking avatars: it highlights separating expressive channels so that behavior can be controlled and audited. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

Thies et al. (2016) present real-time monocular facial capture and reenactment. In this review, the source is used as a governance-and-evaluation comparator for controllable expressivity in three-dimensional talking avatars: it highlights separating identity recovery, expression transfer, and final rendering. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

Prajwal et al. (2020) use a lip-sync expert to guide speech-to-lip generation in unconstrained videos. In this review, the source is used as a governance-and-evaluation comparator for controllable expressivity in three-dimensional talking avatars: it highlights using a task-specific evaluator without confusing synchronization with overall realism. This is a bounded use of the citation. It does not imply that evidence from one domain validates outcomes in another; instead, it makes the design assumption available for explicit review.

5. Implementation Implications and Challenges

Teams applying this synthesis should begin with a compact assurance case. The top claim states the bounded use: what the system is intended to support, for whom, and under which conditions. Subclaims cover input adequacy, mechanism validity, evaluation relevance, operator usability, and governance effectiveness. Each subclaim points to evidence and records remaining uncertainty. This structure makes disagreements productive because reviewers can challenge a specific link rather than accept or reject an entire technology category.

Change management is particularly important. Materials age, environments shift, data distributions evolve, policies are revised, and users adapt to interfaces. A release decision should therefore include monitoring triggers and revalidation criteria. Examples include a change in source material, sensor calibration, label definition, demographic mix, model dependency, moderation policy, or legal purpose. Monitoring is useful only if a threshold leads to an assigned action such as investigation, rollback, retraining, maintenance, or notification.

Procurement and partnership reviews should ask for evidence at the same granularity. Broad claims of accuracy, efficiency, safety, or privacy are insufficient when the operating conditions are unclear. Reviewers should request dataset or material provenance, test protocols, uncertainty reporting, known exclusions, change logs, and incident procedures. Where evidence cannot be shared, the limitation should be treated as unresolved risk rather than filled with assumptions.

6. Discussion

A comparative reading of the sources supports three general observations. First, representations matter: geometry, labels, molecular structure, temporal state, and policy taxonomy all shape what a system can express. Second, controllability matters: temperature, wettability, model routing, generation difficulty, expressive channels, and review thresholds are useful only when operators can observe and adjust them. Third, governance matters: technical outputs acquire consequences through institutional decisions, so auditability and appeal are properties of the complete workflow.

These observations do not create a universal metric. Domain-specific evidence remains indispensable.

A psychometric instrument cannot be validated by a materials study, and a transport exposure study cannot establish the safety of a recommendation model. The value of cross-domain comparison is narrower: it reveals questions that every design team should answer in its own evidentiary language. This limitation is especially important for controllable expressivity in three-dimensional talking avatars, where superficial transfer could magnify rather than reduce uncanny behavior, identity misuse, and aesthetic bias.

The strongest internal position is therefore conditional. A component may be promising for a stated function while the surrounding evidence remains incomplete. Recording that distinction avoids both premature rejection and premature deployment. It also creates a practical research agenda: identify the missing claim, select an aligned method, define an acceptance threshold, and preserve the decision record.

7. Conclusion

This internal review has examined controllable expressivity in three-dimensional talking avatars as part of an evidence-bearing socio-technical or engineering system. The synthesis emphasizes explicit boundaries, mechanism-aware architecture, decision-aligned evaluation, and governance that remains effective as conditions change. For telepresence, digital instruction, and accessible interfaces, readiness should be judged through a traceable chain from inputs and representations to actions, consequences, monitoring, and recovery. The allocated literature provides concrete design comparators, but no cross-domain citation is treated as substitute evidence. The immediate practical recommendation is to build a bounded assurance case before committing to scale, and to keep unresolved assumptions visible until domain-appropriate evidence closes them.

References

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., & Black, M. J. (2019). Capture, learning, and synthesis of 3D speaking styles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10101–10111). IEEE.

Liu, K., Xiong, H., Zhang, J., & Peng, M. (2026). MOSAIC: A Cognitively Motivated Multi-Agent Framework for Interpretable and Training-Free Empathetic Dialogue. Electronics, 15(10), 2078.

Liu, K., Xiong, H., Zhang, J., & Peng, M. (2026). Unifying Aesthetic Evaluation via Multimodal Annotation and Fine-Grained Sentiment Analysis. Big Data and Cognitive Computing, 10, 37.

Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 484–492). Association for Computing Machinery. https://doi.org/10.1145/3394171.3413532

Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., & Sheikh, Y. (2021). MeshTalk: 3D face animation from speech using cross-modality disentanglement. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1173–1182). IEEE.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. ACM Transactions on Graphics, 36(4), Article 95. https://doi.org/10.1145/3072959.3073640

Tao, J., Lyu, R., & Cao, X. (2026). A Deep Learning-Based Automated Content Moderation Framework for Online Platforms. Future-Adaptive Intelligence and Lifelong Systems, 1(1).

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2387–2395). IEEE. https://doi.org/10.1109/CVPR.2016.262

Xiao, F., Song, Y., & Chen, J. (2026, May). DLA-Net: Dynamic Label-Aware Network with Dual-Branch Feature Fusion for Mixed-Type Wafer Map Defect Recognition. In 2026 6th International Conference on Electronics, Circuits and Information Engineering (ECIE) (pp. 72-78). IEEE.

Xiong, H., Zhang, J., Wang, Z., Pan, T., & Hu, Q. (2026). VividTalker: A Modular Framework for Expressive 3D Talking Avatars with Controllable Gaze and Blink. In ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).