



Multimodal Artificial Intelligence for Intelligent Human–Computer Interaction

Marcel Worrying

Professor, University of Amsterdam, Netherlands

Abstract

Human–computer interaction is progressively moving beyond keyboard-, pointer-, and screen-centered communication toward intelligent interfaces capable of interpreting speech, text, facial behavior, gesture, visual context, and other human-generated signals simultaneously. Multimodal artificial intelligence provides the computational foundation for this transition by integrating heterogeneous information sources and allowing interactive systems to infer user intention from complementary evidence. The present study investigates the role of multimodal artificial intelligence in intelligent human–computer interaction through a conceptual analysis and reproducible synthetic simulation. The proposed framework combines speech, visual, and textual observations using quality-aware probabilistic fusion and evaluates interaction accuracy under clean conditions, acoustic noise, visual occlusion, and combined ambiguity. Twenty thousand synthetic binary interaction trials were generated for each environmental condition using a fixed random seed. Speech-only, vision-only, text-only, and adaptive multimodal configurations were compared. The simulation produced multimodal intent-recognition accuracies of 98.49% under clean conditions, 92.91% during acoustic interference, 92.38% during visual occlusion, and 81.28% under combined ambiguity.

The multimodal configuration remained more robust than individual modalities because the fusion mechanism increased the contribution of relatively reliable signals while reducing dependence on degraded channels. The findings demonstrate the potential of multimodal AI to improve interaction robustness, contextual sensitivity, accessibility, and naturalness. Nevertheless, successful deployment requires reliable cross-modal alignment, uncertainty modeling, privacy protection, inclusive training data, efficient inference, and appropriate human control. The study concludes that intelligent HCI should not simply maximize the number of available modalities; rather, it should dynamically integrate those modalities according to their contextual reliability and relevance to the interaction task.

Keywords: multimodal artificial intelligence, human–computer interaction, multimodal fusion, intelligent interfaces, speech recognition, computer vision, natural language processing, adaptive interaction

1. Introduction

Human–computer interaction has undergone a substantial transformation from command-oriented interfaces toward systems intended to interpret human communication in increasingly natural and contextual forms. Conventional interfaces typically require users to adapt their actions to predetermined interaction mechanisms such as keyboards, pointing devices, menus, and structured commands. Human



communication, by contrast, rarely depends on a single channel. People simultaneously use spoken language, facial expressions, gaze, gesture, posture, textual information, and contextual cues to communicate intentions. Multimodal artificial intelligence attempts to reduce this gap by enabling computational systems to process several forms of human information together rather than treating each source independently.

The concept of multimodal interaction predates contemporary deep learning. Oviatt emphasized that multimodal interfaces should not simply be understood as combinations of redundant input channels but as interaction systems in which modalities can complement one another according to users, tasks, and environmental conditions. This principle remains central to modern intelligent HCI. A spoken command may be clear in a quiet room but unreliable in a crowded public environment. A visual gesture may be informative when the user's face and hands are visible but ineffective under poor illumination or occlusion. Text can provide comparatively stable semantic information but may omit emotional tone, urgency, or physical context. An intelligent interface must therefore determine not only what each modality communicates but also how much confidence should be assigned to each source at a particular moment.

Recent advances in machine learning have considerably expanded the technical possibilities for multimodal interaction. Baltrušaitis, Ahuja, and Morency organized multimodal machine learning around key challenges involving representation, translation, alignment, fusion, and co-learning. These challenges are directly relevant to HCI because speech, facial movement, text, gaze, and gesture are produced at different rates, contain different forms of information, and may not align perfectly in time. A system that detects words correctly but associates them with the wrong facial expression or gesture may reach an incorrect interpretation even when every individual modality has been processed successfully.

Multimodality is also relevant to affective and socially responsive computing. Picard's foundational work on affective computing established the importance of computational systems that can recognize and respond to human emotional information. Subsequent research has increasingly combined audio, visual, and linguistic signals because emotions and social states are rarely expressed through a single observable channel. A review by Poria and colleagues documents the shift from unimodal affect analysis toward multimodal fusion, while D'Mello and Kory's meta-analysis found that multimodal affect-detection systems commonly exceeded their strongest unimodal counterparts, although gains depended substantially on data characteristics and evaluation conditions.

Despite these advances, multimodal AI should not be assumed to guarantee intelligent interaction. Additional modalities increase information availability but also introduce synchronization requirements, contradictory signals, computational cost, privacy exposure, missing data, and uncertainty. The present study therefore examines multimodal HCI through the concept of adaptive evidence integration. A reproducible simulation is developed in which speech, visual, and textual channels have different reliability levels under environmental disturbances. An adaptive fusion mechanism combines the modalities according to their estimated reliability. The objective is to examine whether multimodal integration improves interaction robustness and to identify the technical, ethical, and design requirements necessary for responsible intelligent human–computer interaction.

2. Review of Literature

2.1 From Multimodal Interfaces to Multimodal Intelligence

Early research on multimodal interfaces challenged the assumption that interaction would naturally evolve toward one universally superior input mechanism. Oviatt argued that users combine modalities strategically and that multimodal behavior varies with tasks, environments, individual preferences, and communicative demands. This observation is significant because contemporary AI systems sometimes treat multimodality primarily as a machine-learning architecture problem. From an HCI perspective, however, multimodality is fundamentally associated with how humans communicate and adapt their behavior.

Multimodal intelligent interfaces may integrate speech recognition, natural language processing, computer vision, gesture recognition, eye tracking, physiological sensing, touch, and contextual information. These signals are heterogeneous. Speech contains lexical content alongside acoustic characteristics such as pitch, intensity, and rhythm. Images and video contain facial movement, posture, object, and environmental information. Text contributes explicit semantic content but provides comparatively limited nonverbal information. An effective interface therefore requires representations capable of preserving both modality-specific information and cross-modal relationships.

Baltrušaitis, Ahuja, and Morency's multimodal learning taxonomy provides an important theoretical foundation for this requirement. Their framework identifies representation, translation, alignment, fusion, and co-learning as central multimodal challenges. Representation concerns how heterogeneous signals can be encoded meaningfully. Alignment involves identifying relationships among elements across different temporal streams. Fusion concerns combining information for prediction, while co-learning considers how knowledge available in one modality may improve learning in another. Intelligent HCI can involve all five problems simultaneously.

2.2 Fusion, Cross-Modal Attention, and Contextual Interpretation

Multimodal fusion has evolved substantially from simple feature concatenation and decision averaging. Early fusion combines modality-level features before prediction, whereas late fusion combines outputs produced by modality-specific models. More advanced methods explicitly represent interactions among modalities. Zadeh and colleagues introduced the Tensor Fusion Network to model intra-modal and inter-modal dynamics in multimodal sentiment analysis. Such architectures illustrate why the meaning of combined signals cannot always be represented adequately through independent predictions.

Temporal alignment creates an additional challenge. Spoken words, facial expressions, and gestures do not necessarily begin and end at the same moment. A user may smile before speaking, point after issuing part of a command, or change facial expression during a sentence. Tsai and colleagues developed the Multimodal Transformer to address unaligned multimodal language sequences through cross-modal attention rather than requiring rigid pre-alignment. Their work demonstrates the importance of learning relationships between asynchronous linguistic, visual, and acoustic sequences.

These developments have important implications for intelligent HCI. Rather than assuming that every modality is continuously reliable, an interface can estimate signal quality and dynamically alter the contribution of each modality. Acoustic information can receive lower weight during noise, while visual information can be reduced when a face is occluded or poorly illuminated. This concept motivates the adaptive fusion approach examined in the present simulation.

2.3 Affective and Human-Centered Multimodal Interaction

Affective computing extends intelligent HCI beyond literal command recognition toward interpretation of user state. Picard argued that affective information is relevant to intelligent interaction because emotion influences cognition, decision-making, communication, and human responses to technology. This perspective created a foundation for systems that attempt to detect frustration, engagement, stress, interest, confusion, or emotional valence.

Multimodal methods are particularly appropriate for affective interaction because emotional signals are distributed across facial, vocal, linguistic, physiological, and behavioral channels. Poria and colleagues reviewed the transition from unimodal analysis toward multimodal affective fusion. D'Mello and Kory's quantitative review of multimodal affect-detection systems further reported that multimodal systems were frequently more accurate than their best unimodal alternatives, while also showing that performance gains were smaller for natural behavior than for acted expressions. This distinction is important for HCI because laboratory demonstrations can overestimate the reliability achievable during spontaneous real-world interaction.

The literature therefore supports multimodality while simultaneously warning against simplistic performance assumptions. Intelligent interaction requires robust sensing, realistic evaluation, cross-modal modeling, contextual interpretation, and continuous consideration of how users experience the system.

3. Objectives of the Study

3.1 Multimodal Interaction Objective

The primary objective of the study is to examine whether adaptive multimodal artificial intelligence can improve intent recognition when the reliability of individual interaction channels changes across environmental conditions.

The research does not assume that every modality is equally informative. Instead, it investigates an interaction architecture in which the contribution of speech, visual, and textual signals changes according to estimated modality reliability.

The principal objectives are:

- To examine the role of multimodal artificial intelligence in creating more natural and context-aware human-computer interaction.
- To compare unimodal and multimodal intent-recognition performance under clean and degraded interaction environments.
- To evaluate adaptive fusion as a mechanism for reducing dependence on unreliable sensory channels.
- To identify major technical, human-centered, privacy, and governance challenges associated with multimodal interfaces.
- To propose design principles for reliable and responsible multimodal HCI.

3.2 Robustness and Contextual Adaptation

A second objective concerns robustness. An intelligent interface should not perform effectively only when laboratory conditions are ideal. Human interaction occurs in environments containing background speech, changing illumination, partial occlusion, incomplete language, accents, gestures, interruptions, and inconsistent sensor availability.



The study therefore compares four interaction environments: clean conditions, acoustic noise, visual occlusion, and combined ambiguity. These conditions represent controlled abstractions rather than recordings of real users. Their purpose is to examine how complementary modalities might preserve interaction performance when one or more information channels become unreliable.

3.3 Research Questions

The study addresses three related questions. First, does adaptive multimodal fusion outperform individual modalities when interaction conditions vary? Second, does the advantage become more meaningful when a dominant modality becomes degraded? Third, what design limitations prevent simulated multimodal performance from being interpreted automatically as evidence of real-world user experience?

These questions position system accuracy as only one component of intelligent HCI. Practical success also depends on latency, accessibility, user trust, privacy, interpretability, and appropriate failure handling.

4. Research Methodology

4.1 Research Design and Synthetic Interaction Environment

The study follows a simulation-based methodological design. No real participants, audio recordings, facial images, biometric information, or private conversational data were collected. The experiment represents intent recognition as a binary classification problem in which the user's intended command belongs to one of two abstract interaction classes.

For each simulated trial, speech, vision, and text channels independently produce either a correct or incorrect interpretation. Their probabilities of correct interpretation vary across environmental conditions. The clean environment gives all three modalities high reliability. Acoustic noise primarily degrades speech. Visual occlusion primarily degrades vision. Combined ambiguity reduces the reliability of every modality.

4.2 Modality Configuration and Adaptive Fusion

The modality probabilities were selected as controlled experimental parameters rather than empirical estimates.

Table 1: Simulation Configuration for Multimodal HCI

Interaction Condition	Speech Reliability	Vision Reliability	Text Reliability	Primary Disturbance
Clean Environment	0.93	0.91	0.94	Minimal interference
Acoustic Noise	0.64	0.89	0.92	Speech degradation
Visual Occlusion	0.90	0.61	0.91	Visual degradation
Combined Ambiguity	0.68	0.70	0.79	Multiple uncertain channels

4.3 Evaluation Procedure and Reproducibility

Four system configurations are evaluated:

- Speech-only interaction
- Vision-only interaction
- Text-only interaction
- Adaptive multimodal fusion

Recognition accuracy is calculated as:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100$$

The same environmental condition is applied to all configurations. Consequently, the comparison evaluates how the availability and combination of modalities influence simulated intent recognition.

The study deliberately avoids inferential claims concerning real populations. No hypothesis tests, confidence intervals for human performance, or claims about user satisfaction are generated because the observations are synthetic computational trials rather than independent human participants.

The simulation therefore serves as a methodological demonstration of adaptive multimodal integration and not as evidence that a particular deployed HCI system would achieve the same recognition levels.

5. Analysis

5.1 Comparative Recognition Performance

The simulation shows that all modalities perform strongly in the clean environment. Speech-only interaction achieves 93.05% accuracy, vision-only interaction achieves 90.94%, and text-only interaction reaches 94.16%. Adaptive multimodal fusion achieves 98.49%.

This clean-environment result demonstrates one potential benefit of complementary evidence. Even when individual channels are comparatively accurate, each still produces occasional errors. If those errors are not perfectly correlated, information from the remaining modalities can help correct an incorrect interpretation.

The more significant results emerge under degraded conditions. Acoustic noise reduces speech-only recognition to 64.38%, while vision and text remain at 88.87% and 92.17%, respectively. Adaptive multimodal fusion maintains 92.91% accuracy because the degraded speech channel contributes less heavily to the final decision.

Visual occlusion creates the opposite pattern. Vision-only performance declines to 61.47%, whereas speech remains at 89.76% and text at 91.04%. The adaptive multimodal configuration produces 92.38% accuracy.

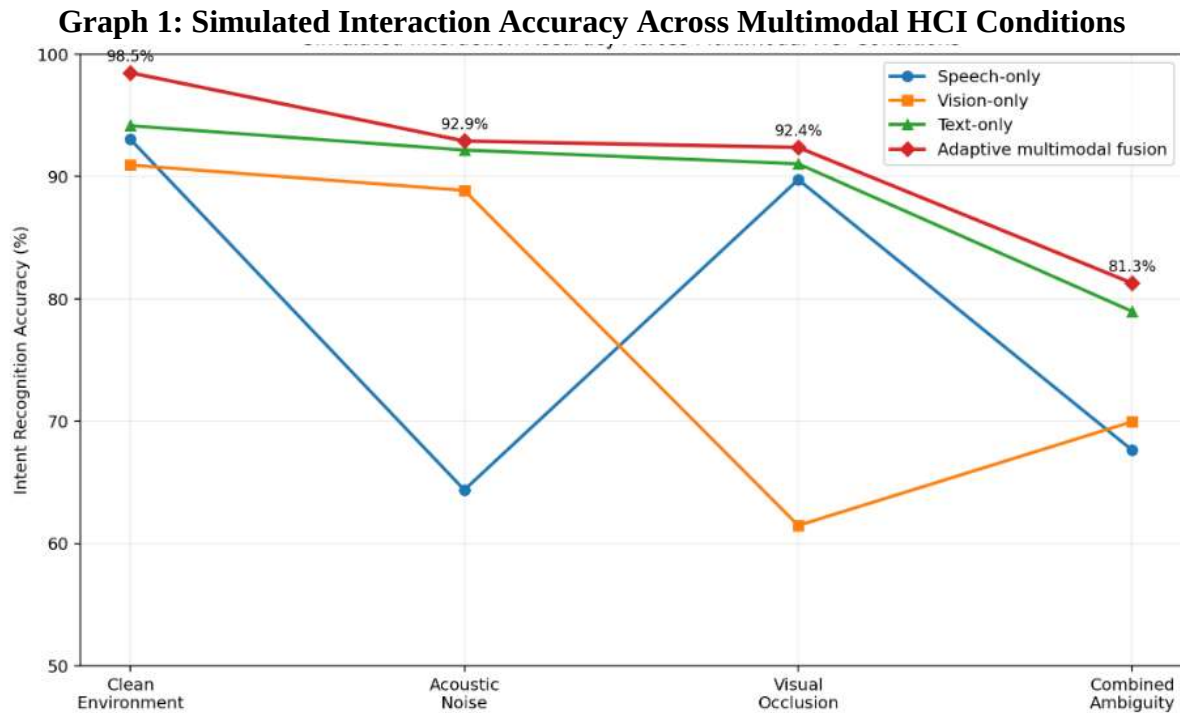
Table 2: Simulated Intent-Recognition Accuracy Across Interaction Conditions

Interaction Condition	Speech-Only	Vision-Only	Text-Only	Adaptive Multimodal Fusion
Clean Environment	93.05%	90.94%	94.16%	98.49%
Acoustic Noise	64.38%	88.87%	92.17%	92.91%
Visual Occlusion	89.76%	61.47%	91.04%	92.38%
Combined Ambiguity	67.62%	69.96%	78.97%	81.28%

Source: Fixed-seed synthetic simulation conducted for the present study.

Note: Values are computational simulation outputs and do not represent measured performance from actual users or a deployed commercial multimodal AI platform.

5.2 Graphical Analysis of Multimodal Robustness



The graphical comparison reveals different degradation patterns among the individual modalities. Speech accuracy declines sharply when acoustic conditions deteriorate, while visual accuracy decreases substantially under occlusion. Text remains comparatively stable because the simulated disturbances have a smaller effect on its semantic information.

Adaptive multimodal fusion records the highest accuracy under every evaluated condition. Its advantage is particularly relevant when the primary sensory channel becomes unreliable. During acoustic interference, multimodal fusion exceeds speech-only recognition by 28.53 percentage points. During visual occlusion, it exceeds vision-only recognition by 30.91 percentage points.

Combined ambiguity produces the most difficult condition. Multimodal accuracy decreases to 81.28%, showing that multimodality cannot completely compensate when all available information sources become uncertain. Nevertheless, the fusion configuration remains above speech-only performance by 13.66 percentage points, vision-only performance by 11.32 points, and text-only performance by 2.31 points.

5.3 Interpretation of Cross-Modal Complementarity

The results illustrate a fundamental principle of multimodal intelligence: its principal advantage does not arise merely from collecting more information. The advantage arises when different information sources have complementary error structures.



If speech and vision fail for exactly the same reasons at exactly the same moments, combining them provides comparatively limited benefit. Multimodal fusion becomes valuable when one channel remains informative while another is degraded.

This principle explains why reliability estimation is essential. A system that averages all modalities equally could allow an unreliable camera or noisy microphone to reduce the quality of an otherwise correct decision. Intelligent fusion should therefore estimate modality quality before assigning influence.

The results are consistent with the wider multimodal-learning literature, which treats fusion, alignment, and representation as substantive modeling challenges rather than simple data aggregation.

6. Challenges

6.1 Cross-Modal Alignment and Conflicting Evidence

Human-generated signals are not perfectly synchronized. A person's gesture may begin before speech, facial expression may change while an utterance is still being produced, and textual information may represent only part of the underlying intention. Temporal misalignment can therefore cause a multimodal system to integrate signals belonging to different moments.

Cross-modal attention provides one possible solution because it allows information from one modality to attend selectively to relevant components of another. The Multimodal Transformer demonstrates how such relationships can be learned without requiring exact pre-alignment of every sequence. Nevertheless, attention mechanisms do not eliminate the need for careful training and evaluation.

Conflicting evidence creates a related problem. A user may verbally state that everything is satisfactory while facial behavior indicates frustration. A gesture may contradict spoken direction. Intelligent HCI cannot always resolve such contradictions through statistical confidence alone because human communication includes irony, politeness, hesitation, ambiguity, and culturally dependent expression. Systems should therefore retain uncertainty rather than forcing every interaction into an overconfident prediction.

6.2 Privacy, Surveillance, and Sensitive Human Signals

Multimodal interfaces potentially collect substantially more personal information than conventional interfaces. Audio can contain conversations involving bystanders. Cameras can capture faces, homes, workplaces, documents, and physical behavior. Physiological sensors may reveal health-related or emotional information. Gaze and behavioral traces can expose patterns users did not intentionally communicate.

Privacy risk therefore increases as interfaces become more perceptive. The design principle should not be to collect every technically measurable signal. Systems should collect only modalities genuinely necessary for the interaction purpose.

On-device processing, data minimization, limited retention, explicit permission, sensor indicators, secure storage, and user-controlled modality activation can reduce risk. Users should also be able to use alternative interaction methods where practical rather than being forced to provide facial, vocal, or behavioral information. A system should not interpret technical capacity to infer emotion or identity as automatic justification for performing that inference.

6.3 Accessibility, Bias, and Human Diversity

Multimodal interfaces are frequently presented as inherently accessible because users can interact through alternative channels. This potential is genuine, but accessibility depends on design quality.

Speech recognition may perform inconsistently across accents, speech impairments, age groups, or background environments. Gesture recognition may assume forms of physical movement that some users cannot perform. Facial-analysis systems may behave differently across demographic groups, lighting conditions, assistive equipment, cultural expression, or atypical facial movement.

A multimodal system can reduce some of these limitations when it permits users to select alternative channels. It can also amplify discrimination if several biased classifiers are combined.

Inclusive training data, subgroup evaluation, accessibility testing, personalization, modality choice, and meaningful fallback mechanisms should therefore be integral to multimodal HCI design rather than treated as later modifications.

7. Discussion

7.1 Multimodality as Adaptive Evidence Integration

The simulation suggests that multimodal AI should be understood principally as a mechanism for integrating uncertain evidence rather than simply as an interface containing several sensors. This distinction is important because an intelligent interface can contain speech, cameras, gesture recognition, and text processing while still making poor decisions if those sources are combined incorrectly.

The clean-condition result illustrates redundancy. Fusion reaches 98.49%, exceeding all individual channels. Under acoustic and visual disturbance, the more important mechanism is complementarity. The system shifts reliance toward modalities that remain comparatively reliable.

This interpretation aligns with the broader multimodal-machine-learning taxonomy proposed by Baltrušaitis and colleagues. Representation, alignment, and fusion determine how information becomes meaningful across modalities, while co-learning provides opportunities for stronger modalities to support weaker or incomplete channels.

Future intelligent interfaces should therefore estimate signal confidence continuously. A speech recognizer should not influence a decision equally when microphone quality deteriorates. A vision model should not retain full authority when the user leaves the camera frame. Adaptive weighting provides one practical basis for such context-sensitive interaction.

7.2 Toward Natural and Affective Human–Computer Interaction

Multimodal AI also changes the conceptual relationship between people and computing systems. Conventional interfaces typically require explicit commands. Intelligent multimodal systems can potentially infer intention from several simultaneous signals and respond before the user formulates a formal machine-oriented command.

A learning system could combine spoken explanation, gaze, and cursor behavior to infer which concept a learner finds difficult. An assistive system could integrate speech and gesture when one communication channel becomes limited. A conversational agent could use linguistic and acoustic information to identify uncertainty in a user's request. These possibilities represent a movement from command execution toward context-sensitive interaction.

Affective information adds further complexity. Research in affective computing has shown substantial interest in combining facial, vocal, textual, and physiological evidence. Yet emotion

inference should remain cautious. Human emotional states cannot always be reliably inferred from visible expressions, and systems should not present probabilistic affect predictions as objective psychological facts. Intelligent HCI therefore requires epistemic restraint. A system can use multimodal signals to adapt support while remaining transparent about uncertainty.

7.3 Reliability, User Control, and Responsible Intelligence

The combined-ambiguity condition provides an important counterpoint to optimistic interpretations of multimodal AI. When every input channel deteriorates, multimodal fusion also deteriorates. The simulation falls from 98.49% in clean conditions to 81.28% under combined ambiguity.

The correct system response in such situations should not always be to make the best available prediction. Intelligent interaction should sometimes recognize that confidence is insufficient and request clarification.

For example, a system could ask a user to repeat a command, confirm an interpreted action, switch modality, or provide explicit selection before a consequential operation is performed. Such behavior can be more intelligent than silently making an uncertain prediction.

Human control is therefore an essential characteristic of responsible multimodal HCI. Users should understand which sensors are active, be able to deactivate nonessential modalities, correct inaccurate interpretations, and override system behavior. The future of intelligent interaction should consequently combine high-quality multimodal perception with transparent uncertainty management and user agency.

8. Conclusion

Multimodal artificial intelligence provides a significant technological pathway toward more natural, robust, and context-sensitive human-computer interaction. Human communication is inherently distributed across linguistic, acoustic, visual, behavioral, and contextual channels, while conventional interfaces frequently capture only a small proportion of this information. By integrating complementary modalities, intelligent systems can reduce dependence on any single interaction channel and adapt more effectively to changing environmental conditions.

The simulation presented in this study demonstrates this principle through speech, vision, text, and reliability-aware multimodal fusion. Under clean conditions, multimodal fusion achieved a simulated intent-recognition accuracy of 98.49%. More importantly, the approach remained comparatively robust when individual modalities deteriorated. Under acoustic noise it retained 92.91% accuracy, and under visual occlusion it retained 92.38%. When all modalities became more ambiguous, performance declined to 81.28%, demonstrating that multimodal integration improves robustness but cannot eliminate uncertainty.

The analysis therefore supports an adaptive rather than additive interpretation of multimodal intelligence. Effective HCI does not require every available sensor to contribute equally at every moment. It requires the system to determine which information is currently reliable, align relevant signals, integrate complementary evidence, detect conflicting inputs, and recognize situations in which uncertainty remains too high for autonomous interpretation.

Future research should extend the present synthetic model to real multimodal datasets, spontaneous user interactions, multilingual speech, culturally diverse participants, accessibility scenarios, missing modalities, cross-modal transformers, and longitudinal user studies. Evaluation

should include not only recognition accuracy but also latency, cognitive workload, task completion, accessibility, privacy perception, trust, fairness, and user satisfaction. The long-term value of multimodal AI will depend not merely on whether machines can perceive more human signals, but on whether those signals are interpreted responsibly and used to create interaction that remains understandable, controllable, inclusive, and genuinely supportive of human goals.

References

1. Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, 42(11), 74–81. <https://doi.org/10.1145/319382.319398>
2. Oviatt, S. (2007). Multimodal interfaces. In J. A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook* (2nd ed., pp. 413–432). Lawrence Erlbaum Associates.
3. Dumas, B., Lalanne, D., & Ingold, R. (2009). Pattern recognition techniques applied to gesture recognition: A review. In *Proceedings of the 8th International Conference on Automatic Analysis of Human Behavior*.
4. Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, 36, 189–195. <https://doi.org/10.1016/j.patrec.2013.07.003>
5. Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1–2), 116–134. <https://doi.org/10.1016/j.cviu.2006.10.019>
6. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>
7. Morency, L.-P., Mihalcea, R., & De Mori, R. (2005). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th ACM International Conference on Multimedia*, 299–302. <https://doi.org/10.1145/1101149.1101211>
8. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing research based on function-component-context framework. *IEEE Transactions on Affective Computing*, 8(2), 144–162. <https://doi.org/10.1109/TAFFC.2016.2584327>
9. Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016*, 1601–1612.
10. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. <https://doi.org/10.18653/v1/D17-1115>
11. Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
12. Liang, P. P., Liu, Z., & Morency, L.-P. (2018). Multimodal language analysis with recurrent neural networks. *IEEE Signal Processing Magazine*, 35(5), 108–117.
13. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, 689–696.
14. Srivastava, N., & Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. In *Advances in Neural Information Processing Systems 25*, 2222–2230.
15. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, 689–696.



16. Senechal, T., Ringeval, F., Ros, F., & others. (2017). Multimodal emotion recognition using audiovisual information. *Proceedings of the ACM International Conference on Multimodal Interaction*.
17. Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42, 335–359. <https://doi.org/10.1007/s10579-008-9076-6>
18. Morency, L.-P., Zadeh, A., & Morency, L.-P. (2017). Multimodal sentiment analysis and machine learning. In *Proceedings of the 2017 ACM International Conference on Multimodal Interaction*.
19. McDuff, D., El Kaliouby, R., & Picard, R. W. (2011). Affective computing and intelligent human-computer interaction. *IEEE Signal Processing Magazine*, 28(2), 139–141.
20. Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, 43(4), 70–78. <https://doi.org/10.1145/332051.332075>