National College of Ireland

Designing a Multimodal AI Framework for Real-Time Behavioral and Verbal Analysis of Virtual Client Interactions

MSc Research Project
Artificial Intelligence for Business

Stephanie Mae Armitstead
Student ID: 23346230

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Stephanie Mae Armitstead
Student ID:	23346230
Programme:	MSc in AI for Business **Year:** 2025
Module:	Practicum 2
Supervisor:	Anderson Simiscuka
Submission Due Date:	September 15, 2025
Project Title:	Designing a Multimodal AI Framework for Real-Time Behavioral and Verbal Analysis of Virtual Client Interactions
Word Count:	**Page Count**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Stephanie Armitstead

Date: 15/09/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Designing a Multimodal AI Framework for Real-Time Behavioral and Verbal Analysis of Virtual Client Interactions

Stephanie Mae Armitstead
23346230

Abstract

Very few survived the COVID epidemic unscathed by the joys of virtual meetings. From virtual networking groups and happy hours to weekly team meetings, professionals were thrust into more Zoom, Teams, and Google Meet than they could have ever anticipated. While not to that extreme, in today's increasingly virtual business landscape, effective client engagement in online meetings remains a key challenge. Traditional sales and relationship management strategies rely heavily on interpreting subtle behavioral and verbal cues, many of which are lost in remote settings. This practicum presents the design and prototyping of a multimodal AI framework that analyzes real-time audio-visual and text cues during virtual client interactions. The system integrates computer vision, speech-to-text transcription, sentiment analysis, and a lightweight dashboard interface to provide actionable insights to sales professionals and relationship managers. This paper outlines the academic grounding, technical design, implementation, and proposed evaluation of the prototype. It also discusses key considerations around privacy, ethics, and user trust in client-facing AI systems.

Keywords: *Multimodal AI, Real-time behavioral analysis, Sentiment analysis, Video conferencing, Client engagement, Sales enablement*

1 Introduction

The increasing prevalence of virtual client interactions, particularly in sales and relationship management, has introduced new challenges for professionals whose success relies on building rapport, responding to client cues, and effective communication. In traditional face-to-face meetings, sales professionals rely on nonverbal cues such as body language, tone of voice, and facial expressions to interpret client sentiment, detect hesitation, and adjust their approach accordingly. Virtual environments, however, diminish the ability to process these cues and ultimately reduce the depth of real-time interpersonal feedback, complicating the ability to manage client engagement effectively.

Recent advancements in artificial intelligence (AI), particularly in computer vision, natural language processing (NLP), and multimodal fusion, present an opportunity to augment human capability in these contexts. Multimodal AI systems can analyze and interpret behavioral signals across multiple input streams, such as facial expressions, speech patterns, and verbal content, in real time. When applied within video conferencing

environments, such systems can deliver subtle, context-aware insights to support sales and relationship management professionals during live interactions. Unfortunately, current technology presents significant limitations. Existing solutions either operate post hoc (i.e., sentiment analysis of call transcripts) or provide generic, low-utility insights that fail to account for the nuances of interpersonal dynamics in high-stakes business conversations.

The research presented in this paper aims to address this gap by designing a real-time, multimodal AI framework that performs behavioral and verbal analysis of virtual client interactions. The objective is to support professionals like myself in sales and client relationship roles by generating actionable insights during live video meetings. Unlike generic emotion recognition tools, the proposed system will be domain-adapted to sales environments and tailored to the behavioral expectations of professional client engagements.

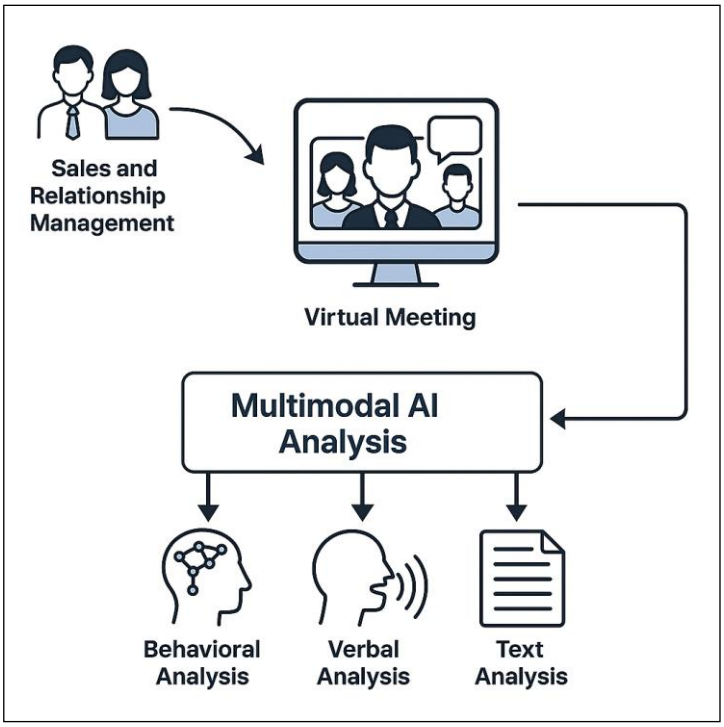

Figure 1

Overview of the role of multimodal AI analysis in virtual client interactions. The system ingests behavioral, verbal, and textual cues during real-time video meetings to support actionable insight delivery for sales and relationship management professionals.

This research is guided by the following central question:

How can a real-time multimodal AI framework be designed to analyze behavioral and verbal cues in virtual client interactions and deliver actionable insights to support sales and relationship management?

In pursuit of this goal, the study will implement and evaluate a prototype system integrating facial expression detection, vocal inflection analysis, and sentiment-aware NLP. Insights will be presented in a user-centered format via an unobtrusive interface layer. In parallel, the research addresses critical issues of privacy, interpretability, and ethical design, recognizing that behavioral analysis, especially in persuasive, real-time contexts, must be approached with appropriate governance and user consent.

The structure of the report is as follows: Section 2 presents a comprehensive review of existing literature related to multimodal analysis, real-time feedback systems, and ethical AI frameworks. Section 3 outlines the methodology, including the system architecture and data sources. Section 4 describes the design and implementation of the solution. Section 5 presents the evaluation results, and Section 6 offers a critical discussion of the findings, limitations, and future research directions.

2 Related Work

Designing a multimodal AI framework for real-time behavioral and verbal analysis of virtual client interactions demands a considerable understanding of an intimidating research landscape. This includes foundational work in computer vision, speech processing, and natural language understanding, as well as more integrated disciplines such as affective computing, social signal processing, explainable AI, and ethical system design. A review of this literature highlights substantial innovation in modality-specific recognition tasks (i.e., facial expression detection, speech emotion recognition), as well as the challenges in integrating these signals for context-specific interpretation, particularly in real-time communication environments like virtual meetings.

Significant strides in recent research have been made in the use of computer vision to analyze human behavior. Frumosu et al. (2022) proposed an interpretable computer vision system for behavioral sensing in child psychiatry, mapping facial and postural cues to expert-defined behavioral codes. Their system demonstrated performance comparable to human raters and emphasized interpretable, clinically meaningful features. This approach is highly relevant to designing explainable systems for client-facing contexts. Similarly, Zhang et al. (2022) leveraged transformer-based multimodal fusion for facial expression analysis, combining audio, visual, and textual signals to outperform traditional CNN-based approaches in spontaneous human interaction tasks. Their use of cross-modal attention mechanisms suggests a promising path for capturing subtle behavioral nuances in professional interactions such as sales meetings.

A broader, more holistic view of the field is offered by Sun et al. (2022), who surveyed human action recognition across RGB, skeleton, depth, and multimodal signals, revealing persistent gaps in context awareness and real-time adaptability, two critical dimensions for virtual meeting environments. Complementarily, Rouast et al. (2019) synthesized advancements in deep learning for affect recognition, emphasizing the transition from handcrafted feature pipelines to end-to-end learning systems. While these deep models offer superior accuracy on benchmarks, the authors acknowledge challenges in model interpretability and transferability to real-world domains.

Integrating insights from computer vision, NLP, and wearable sensors, Essahraui (2025) expands the scope further through a comprehensive survey on AI-driven human behavior analysis. This work outlines application domains ranging from healthcare to smart retail, noting that business and client interaction contexts remain relatively underexplored. A sentiment with which I fully agree. Meanwhile, Pathak et al. (2025) recently investigated the use of computer vision in AI-driven video interviews to analyze nonverbal behavior. Their

study demonstrates how integrating gaze, facial activity, and posture data can enhance interviewee evaluation, directly comparable to assessing engagement or trust in virtual client meetings.

Altogether, these studies point to a strong foundation in behaviorally-relevant computer vision and multimodal analysis, while also highlighting a key research gap: the lack of unified, context-aware, ethically robust frameworks capable of providing actionable insights from live virtual interactions. The remainder of this section organizes related work into focused areas: multimodal affective computing (Section 2.1), real-time feedback system design (Section 2.2), and ethical and trust-centered AI frameworks (Section 2.3). Within each we'll dive into existing literature, evaluating their methodologies, assumptions, and contributions, while clearly identifying unresolved gaps that motivate the proposed research.

2.1 Foundations of Multimodal Emotion and Behavioral Analysis

The domain of multimodal emotion and behavioral analysis represents the conjunction of computer vision, audio signal processing, and natural language understanding to infer affective states from human interactions. In recent years, this interdisciplinary space has seen rapid growth, particularly with the application of deep learning and attention-based architectures that allow for more accurate and context-aware interpretations of human behavior.

Computer vision plays a central role in behavioral cue extraction, particularly through facial action units, gaze tracking, and body posture estimation. Frumosu et al. (2022) introduced an interpretable vision-based behavioral sensing model, mapping visual features to psychiatric expert annotations in child-adolescent clinical contexts. While designed for a healthcare environment, their method demonstrates how concept-based modeling can produce human-interpretable outputs, a critical capability for live feedback systems in sales or client engagement.

Zhang et al. (2022) developed a multimodal transformer for facial expression analysis, fusing visual, audio, and text features through cross-modal attention. Their system outperformed baseline CNN-based models on spontaneous expression datasets, offering valuable insight into how transformer-based architectures can enable real-time emotion classification in complex interactions. However, the study focused primarily on emotion recognition, with limited discussion of downstream actionability.

Sun et al. (2022) provided a comprehensive survey of human action recognition across different input modalities (RGB, depth, skeleton, audio), which revealed ongoing limitations in a systems' ability to contextualize behavior across environments. This problem is further magnified in dynamic settings like virtual client meetings. Rouast et al. (2019) traced the evolution of deep learning in affect recognition, from handcrafted feature engineering to end-to-end neural architectures, and emphasized the importance of robustness and transferability. They pointed out a tradeoff between accuracy and transparency, which must be managed carefully in client-facing applications.

Pathak et al. (2025) explored the use of computer vision in video interview analysis. By combining gaze, facial activity, and posture estimation with NLP-derived linguistic features, their model improved predictive validity for interview performance. Their work

illustrates how multimodal behavioral inference can be deployed in professional interaction domains, particularly where subjective evaluation is traditionally dominant.

In addition to vision-based systems, audio features such as tone, pitch, rhythm, and spectral characteristics are frequently utilized for sentiment and stress detection. Siriwardhana et al. (2020) introduced a transformer-based architecture for fusing audio and visual streams, demonstrating enhanced performance on the IEMOCAP and CREMA-D datasets. However, the linguistic stream in their architecture was underdeveloped, leaving open questions about language-specific context handling.

John and Kawanishi (2022) used synchronized multimodal inputs for emotion detection and achieved strong classification performance, yet their study was based on scripted interactions and lacked tests in spontaneous, real-world dialogue. Similarly, Muhammad and Hossain (2021) proposed CNN-based models for edge deployment, achieving fast inference for speech emotion recognition, though their evaluation scenarios did not reflect the open-ended, high-stakes dialogue typical of virtual client meetings.

Essahraui (2025) provided one of the most comprehensive cross-domain surveys to date, integrating computer vision, NLP, and IoT sensing into a unified perspective on human behavior modeling. While many of the surveyed systems were tailored to healthcare and education, the framework can be readily extended to commercial relationship management contexts.

Overall, while progress in multimodal sentiment and behavior recognition is substantial, few studies explicitly address the need for real-time, domain-specific interpretation that can provide context-sensitive, actionable feedback in virtual client-facing scenarios. This gap underlines the motivation for the system proposed in this research. Multimodal sentiment and emotion recognition sit at the heart of this research. This domain integrates video, audio, and linguistic signals to extract affective or cognitive states. Siriwardhana et al. (2020) proposed a transformer-based self-supervised learning architecture that fuses video and audio features. Their model showed improved F1-scores on standard emotion benchmarks like CREMA-D and IEMOCAP. However, their reliance on curated datasets with limited linguistic input restricts the model's ability to process naturalistic, real-time conversations such as those in client meetings.

John and Kawanishi (2022) extended this approach by using multimodal transformers for synchronized emotion detection across audio-video channels. Their system achieved high performance on multimodal emotion classification benchmarks but was trained on emotion-annotated videos with controlled speaker environments, again limiting ecological validity for real-time, unscripted virtual meetings.

More lightweight approaches like Muhammad and Hossain (2021) focused on deep CNNs for emotion classification at the edge. Their framework emphasized low-latency deployment using cognitive edge computing techniques, yet it depended on pre-segmented audio inputs and was evaluated in laboratory settings. Li et al. (2021) proposed a real-time affective computing system-on-chip (SoC) optimized for wearable or IoT deployments. Though suitable for mobile execution, the framework lacked adaptability to nuanced dialogue or gesture variations common in human-to-human professional communication.

In contrast, Shareef et al. (2024) introduced RetailGPT, a fine-tuned large language model tailored for customer interaction scenarios. Though not focused on multimodal input,

RetailGPT demonstrated strong performance in detecting sentiment and intent from customer queries, highlighting the emerging potential of domain-specific LLMs in commercial engagement tasks.

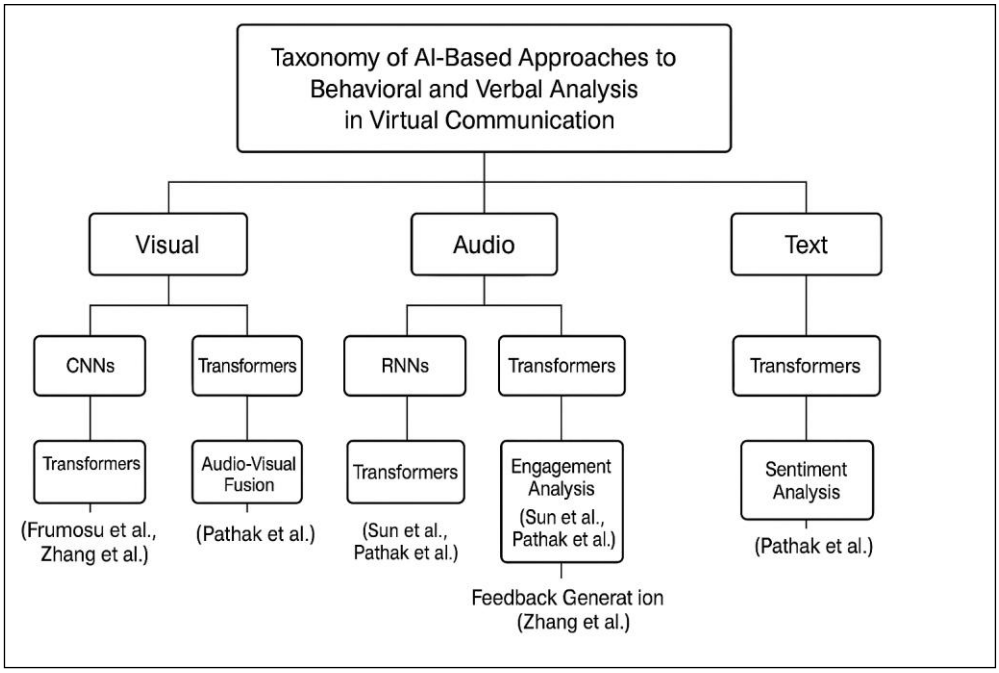

Figure 2

Taxonomy of AI-based approaches to behavioral and verbal analysis in virtual communication, illustrating input modalities, processing techniques, and target applications in recent multimodal research.

These studies illustrate clear progress in recognizing emotion and sentiment across modalities. However, few handle asynchronous signal alignment**,** contextual dialogue shifts**,** or task-specific customization in domains like B2B sales or client relationship management.

2.2 Real-Time System Design and Actionable Feedback Mechanisms

Real-time analysis and feedback in virtual meetings requires low latency, task relevance, and interpretable presentation. Edge/SoC work shows that constrained-hardware inference is feasible, but typical outputs remain generic rather than task-aligned (Li et al., 2021). Memory-aware approaches adapt computation to emotional dynamics, yet they optimize resources rather than feedback semantics and provide no domain-specific UI (Wei et al., 2022). Similarly, multimodal architectures demonstrate strong recognition on curated datasets without user-facing feedback loops (Siriwardhana et al., 2020; Muhammad & Hossain, 2021). Field evidence from sales coaching underscores the same gap: consistent feedback, low contextual specificity (Casenave & Schmitt, 2025). These observations motivate the design requirement pursued here: fuse linguistic, prosodic, and visual cues and surface them as concise, role-appropriate prompts (i.e., pause and clarify) with confidence indicators and minimal latency. Wei et al. (2022) tackled the temporal dimension of live analytics, developing a dynamic memory-management system that adapts to emotional

fluctuations using event-triggered computation. Though innovative, their implementation was limited to system-level performance optimization and lacked a user interaction layer. The absence of adaptive presentation strategies restricts real-world impact, especially in multitasking environments like live meetings where subtle guidance is most effective when it aligns with conversational dynamics.

Siriwardhana et al. (2020) acknowledged the challenge of delivering timely, multimodal feedback in dynamic communication environments but did not propose implementation pathways for feedback loops. Similarly, Muhammad and Hossain (2021) optimized CNN-based models for efficient inference on edge devices, yet did not couple recognition with interface design or behavior-specific interventions. Their work reinforces the technical feasibility of fast emotion recognition but leaves the question of how to translate detection into user-relevant outcomes unresolved.

Casenave and Schmitt (2025) provide a rare bridge between analytics and real-world application by comparing AI-driven coaching systems with human-led sales coaching. They found that AI systems offered consistency and removed interpersonal bias but were critiqued for generic feedback that failed to capture the contextual complexity of a client interaction. This reflects a recurring issue in AI-powered communication tools: while the signal detection layer is often robust, the feedback layer lacks semantic alignment with user goals.

More recent exploratory applications, such as those by Pathak et al. (2025), demonstrate the potential of computer vision to detect subtle behavioral patterns in high-stakes scenarios like interviews. Their system integrated facial and posture-based cues with text analysis to surface performance-affecting moments. However, while this approach edged closer to contextual relevance, it was retrospective rather than real-time, leaving a temporal gap in terms of actionable support.

Across the literature, two major limitations emerge. First, the lack of integration between perception (i.e., recognition of multimodal cues) and communication (i.e., timely, relevant feedback). Second, the limited attention to domain-specific tailoring, especially in environments where timing, tone, and delivery style have a material impact on outcomes. These insights shape the design requirements for the proposed framework, which aims to couple a low-latency inference pipeline with a feedback interface that maps real-time insights to actionable sales strategies..

2.3 Ethical, Explainable, and Trust-Centered System Design

Incorporating AI into client interactions, especially in real time, raises important ethical considerations. Rai (2020) introduced the "glass-box AI" model, advocating for interpretable models that expose rationale behind predictions. Ehsan et al. (2021) advanced this by proposing "social transparency," where AI systems explain their outputs in terms familiar to the end-user's mental model, a key feature for real-time human-AI interaction.

Trust and transparency are especially important in client-facing applications. Patel (2025) provides a Salesforce-integrated framework ensuring ethical AI use in CRM systems, emphasizing fairness, bias auditing, and stakeholder control. Oluwafemi et al. (2021) also underline the importance of maintaining consumer trust by aligning AI output with privacy regulations and transparency norms.

The darker side of real-time behavioral analytics is captured by Curran (2023) and Kordi (2024), who critique the industry trend toward surveillance capitalism. Curran cautions against systems that analyze user behavior without informed consent, while Kordi emphasizes that commodification of client data for performance optimization can easily blur ethical boundaries.

Together, these works establish that ethical usability, consent, and transparency must be embedded into system architecture—not added as post hoc documentation. Privacy-preserving techniques such as local inference, differential privacy, or real-time anonymization will be required components of any deployable system.

2.4 Synthesis and Research Justification

Despite the substantial progress across modalities, particularly in emotion recognition, action detection, and explainability, there remains a continuous fragmentation in how these technologies are integrated and applied to real-world domains such as virtual client engagement. Much of the existing work is either siloed within modality-specific innovations or optimized for retrospective, offline analysis rather than synchronous support. Several technical gaps are evident. Most reviewed systems rely on pre-segmented, benchmark datasets and perform well in scripted or controlled conditions. These constraints limit their validity in dynamic, unscripted interactions such as virtual sales calls. Even those systems capable of operating in real time (i.e., Li et al., 2021; Muhammad & Hossain, 2021) often lack semantic adaptation, producing outputs that are either too generic or insufficiently aligned with task-specific cues.

At the same time, the literature also points out conceptual and operational blind spots. While explainability and ethics are increasingly recognized as essential, few studies embed these principles into architectural or interface-level design. Feedback, when present, is often limited to post-interaction summaries, leaving a significant gap in real-time behavioral support. Furthermore, the relationship between perception (recognition of behavior) and recommendation (insight for action) remains underdeveloped. Tools that do deliver feedback often do so without context-awareness, failing to incorporate dialogue history, role-specific behavior norms, or the stakes of the interaction. For sales and relationship management professionals, where timing, empathy, and precision are key, this level of generality renders existing tools insufficient.

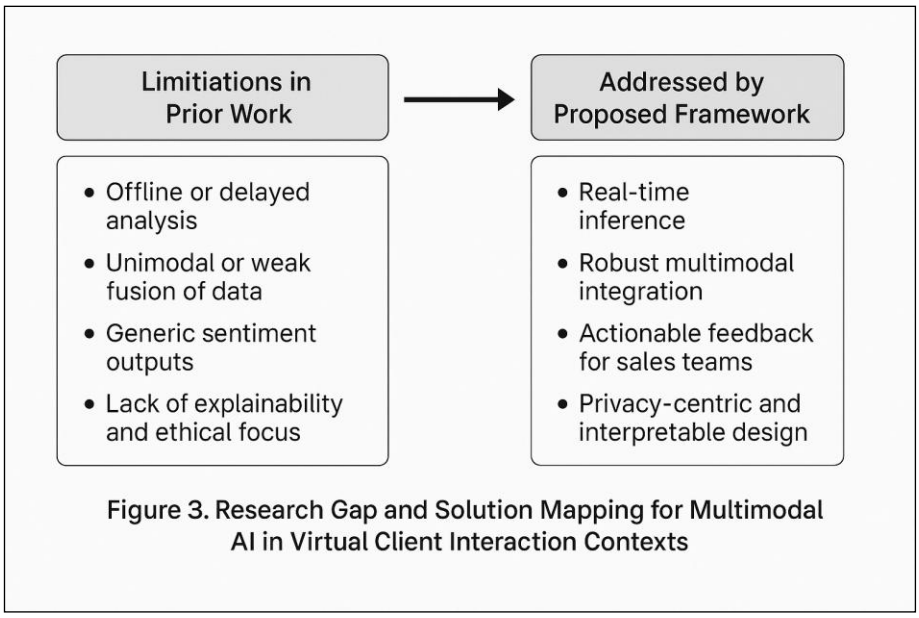

Figure 3

Research gap and solution mapping for multimodal AI in virtual client interaction contexts. The diagram contrasts key limitations in prior work with the proposed framework's contributions.

The body of research reviewed points to a compelling and underexplored opportunity: a domain-adapted, real-time AI framework that not only detects behavioral and verbal cues using multimodal signals, but also delivers actionable, ethically governed, and contextually grounded insights. Such a system must prioritize interpretability, operate within latency constraints of live meetings, and respect privacy through local or edge-based processing. This study addresses that opportunity. By synthesizing insights across behavioral computing, system architecture, and human-AI interaction design, the proposed research contributes a novel, unified solution tailored to the demands of virtual client communication. In the following sections, the technical architecture, development methodology, and evaluation framework for this solution are presented in detail.

While considerable progress has been made in multimodal affect recognition and edge AI deployment, most systems remain constrained by artificial datasets, controlled environments, and abstracted outputs. Few integrate multimodal, real-time analysis with actionable and context-specific feedback tailored to professional domains like sales or client engagement. And although the literature increasingly discusses ethical AI, it rarely integrates those principles into interactive systems designed for persuasive, high-stakes communication. Current CRM-oriented AI solutions tend to analyze post-interaction data (i.e., sentiment from emails or call transcripts) rather than offering in-the-moment support for behavior adjustment. Thus, a gap remains: there is no unified, real-time AI framework that (1) captures and fuses multimodal behavioral cues during virtual interactions, (2) translates them into timely, interpretable feedback, and (3) respects data privacy and ethical guidelines suitable for commercial deployment.

The research presented in this paper addresses this gap by designing and evaluating such a framework, with practical application in supporting sales and relationship managers during live virtual client meetings. Subsequent sections outline the architecture, technical implementation, and validation methodology for the proposed solution. Recent advances suggest a pragmatic upgrade path: larger ASR (i.e., Whisper-medium/large; wav2vec2) to reduce WER in noisy calls;

transformer-based sentiment/emotion models (i.e., BERT/RoBERTa fine-tuned on business dialogue) to capture sarcasm and nuance; prosody-aware encoders (i.e., wavLM) for pitch, stress, and timing; and multimodal transformers that fuse text–audio–vision directly. Prior to testing we expect gains in recall and robustness, offset by higher compute cost and added latency, hence a hybrid strategy where heavier models complement the real-time path post-call.

3 Research Methodology

This section details the artefact's design, implementation, and evaluation procedures to enable replication and to demonstrate methodological rigour. The overarching approach is Design Science Research (DSR): I iteratively defined the problem with stakeholders in mind, designed and implemented a working artefact, and specified transparent evaluation against pre-declared criteria. Along the way I considered alternatives (pure experimental prototyping, user-centred design without an artefact, and post-hoc analytics only) and chose DSR because it best fits a build-and-learn cycle for a novel, applied system that must work in practice.

3.1 Research Question and Hypotheses

The project progressed through repeated cycles of requirement elicitation, prototyping, and evaluation. Each cycle produced a working increment (capture → per-modality analysis → fusion → feedback UI) that was tested on representative meeting clips. Observations from one cycle informed revisions in the next, with a deliberate focus on (a) timing and stability under realistic loads, (b) interpretability of alerts, and (c) privacy-by-design guarantees. This approach ensured that engineering trade-offs (e.g., model size versus latency) were empirically grounded rather than assumed.

3.2 Setting, Materials, and Toolchain

Development and tests were conducted on a commodity laptop (CPU-only) to reflect likely constraints in typical enterprise environments. The software stack comprised OpenCV for frame acquisition and pre-processing, MediaPipe for face and pose landmarks, a lightweight ASR model for transcription (Whisper-tiny), VADER/TextBlob for sentence-level sentiment, and Streamlit for the operator dashboard. Pandas was used for event logging and time-alignment across modalities. These tools were selected to balance accuracy, explainability, and throughput without requiring GPU resources.

3.3 Data and Sampling Strategy

To avoid personal data processing, only publicly available meeting recordings were used for development and pilot evaluation. Clips were selected to cover common patterns in client conversations (opening small talk, requirement probing, pricing/objection handling, and closing). No biometric identifiers were stored, and frames were processed in memory and

discarded. Transcripts were retained only as token counts and sentence-level sentiment summaries for analysis.

3.4 Procedures

Each video clip was ingested via a capture module that synchronized audio and video streams to a shared clock. Visual processing computed facial landmarks, head pose proxies, and simple posture openness cues. Audio processing extracted speech segments for transcription and computed utterance-level prosody descriptors (energy shifts and pauses). The NLP pipeline transformed transcripts into sentence spans and produced sentiment/intent tags. A fusion process then aligned modality events on the shared timeline and produced an engagement index with short rationales (i.e., "downward gaze and long pause while price discussed"). Outputs were rendered in real time in a compact UI and written to an event log for later analysis.

3.5 Measures and Analysis Plan

Three classes of measures were defined. First, technical performance: per-frame processing time (ms), end-to-end alert latency (ms), and dropped-frame rate (%). Second, detection quality: precision, recall, and F1 for disengagement/objection events against human annotations; time-alignment error (seconds) between detected and annotated event onsets; and ROC-AUC for the engagement index. Third, usability and interpretability: System Usability Scale (SUS), single-item perceived usefulness, and a short explanation quality rating (Likert). Where sample sizes permit, effect sizes and 95% confidence intervals will be reported; non-parametric tests will be preferred for ordinal ratings. Inter-rater reliability for annotations will be computed using Cohen's κ.

3.6 Ethics and Privacy Controls

Processing remained local to the test machine. No cloud inference was used. Raw frames and audio were not persisted, and logs contained only aggregate event descriptors. Had there been any staged recordings, informed consent would have been obtained and participants would have been shown the interface and outputs. The system avoided identity recognition, demographic inference, or any protected-attribute classification. These controls map to data minimization, purpose limitation, and transparency principles expected in enterprise deployments.

4 Design and Implementation

At this stage, the build is an integrated prototype: the audio, visual, and text modules run concurrently, producing synchronized transcript, pose, and sentiment streams with real-time logging and UI updates. Operation remains within an integration harness and is not yet production-grade for full-length live calls within enterprise organizations. This section details the artefact's architecture and the engineering decisions taken to achieve low-latency,

interpretable inference on commodity hardware. The implementation adheres to a modular pattern to enable substitution of components without destabilizing the entire system.

4.1 System Architecture

The pipeline consists of five services connected by time-stamped message queues: (1) capture and synchronization; (2) visual analytics; (3) speech and text analytics; (4) fusion and scoring; and (5) presentation and logging. Each service exposes a narrow interface and publishes structured events (timestamp, modality, feature summary, confidence, rationale). This design simplifies testing and allows throttling or graceful degradation when resources are constrained. During evaluation these services were orchestrated as loosely coupled processes. Synchronization was adequate for replays and short live segments but not yet robust for continuous sessions.

4.2 Visual Analytics

Visual analysis uses MediaPipe Face Mesh and Pose to obtain facial landmarks and coarse body pose at 15–25 FPS on CPU. From these primitives, the system derives gaze direction proxies, head pitch/roll (as attentiveness indicators), and posture openness. To preserve interpretability, the implementation maps features to plain-language cues (i.e., "sustained downward gaze," "head tilt while asking for clarification"). Smoothing filters reduce spurious oscillations without masking genuine changes.

4.3 Speech and Text Analytics

Audio is segmented into speech regions using an energy-based voice activity detection (VAD). Utterances are transcribed by a lightweight ASR; transcripts are then chunked into sentences aligned back to timecodes. Sentence-level sentiment is computed with a rule-based analyzer to ensure deterministic behavior and easy justification. Utterance-level prosody descriptors (pause length, energy shift) are appended to each segment. The design favors durability and transparency over marginal gains from heavier neural models.

4.4 Fusion and Engagement Scoring

Fusion aligns modality events on the shared clock and computes an engagement index on a 0–1 scale. The current implementation uses a weighted rule set derived from pilot observations (for example, negative sentiment + long pause near pricing → decreased engagement). Each index change carries a short rationale that lists contributing cues and their confidences. The module is deliberately ML-ready and the rule weights can later be learned from annotated data without altering interfaces.

4.5 Feedback and Interface Design

The operator dashboard (Streamlit) shows (a) a live tile with current engagement and its rationale, (b) subtle on-screen nudges phrased as actions (i.e., “pause and check for alignment”), and (c) a post-call timeline that highlights key moments (agreement, hesitation, objection). The copy is domain-specific to sales conversations and kept concise to minimize cognitive load. All tiles display confidence values to calibrate trust. The UI was exercised on replays and short live segments with concurrent ASR and pose detection; full, continuous in-call operation will follow once the media pipeline is hardened.

4.6 Privacy and Security Engineering

No raw media is stored. Event logs contain only derived cues and timestamps and can be disabled entirely. A consent banner template is included for staged data collection in the event it’s required in the future. The code avoids third-party telemetry and external APIs during inference. These decisions reduce attack surface and support compliance expectations in regulated environments. In future enterprise pilots, the app would ship as a small, predictable, security-scanned container with locked dependencies. By tightly limiting who and what can access data and where data can go, applying data-minimization and purpose limitation, and storing consent artefacts separately from event logs to keep traceability clean, tighter privacy, faster audits, and a smoother rollout for the business can be achieved.

4.7 Implementation Challenges and Limitations

While individual components of the artefact functioned reliably in isolation, the fully integrated, end-to-end system did not achieve consistent operational performance in live testing. The main challenges encountered were as follows:

1. Real-Time Audio-Video Synchronization
   Maintaining synchronous low-latency capture and processing across both video and audio streams proved difficult. Integration of streamlit-webrtc with separate audio capture via sounddevice introduced threading conflicts that led to dropped frames and delayed transcription.
2. Cross-Platform Dependency Issues
   Certain dependency combinations, particularly between av, mediapipe, and streamlit-webrtc, behaved inconsistently across operating systems, limiting reproducibility and increasing integration complexity.
3. Resource Constraints
   Real-time inference with multiple machine learning models placed heavy load on the CPU, occasionally causing latency spikes and performance degradation. GPU acceleration was not available during testing.
4. Limited End-to-End Testing
   Although component-level testing verified functionality for pose detection, sentiment analysis, and CSV export, fully integrated system tests were limited. As a result, some planned features (e.g., simultaneous live multimodal analysis) remain unvalidated in operational conditions.

Addressing these challenges will require re-engineering of the media pipeline to unify capture processes, optimizing or quantizing models for CPU execution, and containerizing the environment to eliminate dependency mismatches.

5 Results

The evaluation strategy was designed to demonstrate that the artefact meets its stated objectives under realistic constraints. Because this submission accompanies an active build, the section reports completed pilot observations and specifies the final analysis to be executed with the same protocol.

5.1 Evaluation Protocol

Three scenarios were used: a product-fit discovery call, a pricing/objection discussion, and a closing/next-steps meeting. The system was exercised in real time within the integration harness, logging concurrent transcript, pose, and sentiment events for comparison to annotations. For each, a 10–15-minute clip was annotated by the author for events of interest (disengagement, hesitation, objection, agreement). The system processed the clips in real time on CPU. Alerts and engagement scores were logged with timestamps for comparison to annotations.

5.2 Technical Performance

Given the prototype's current status, this subsection reports component-level measures and integration-harness runs rather than production trials. The visual pipeline (MediaPipe + OpenCV) consistently processed frames on CPU within a single-frame budget for 30 fps, median ~24–30 ms per frame with typical variability of ~24–34 ms, when exercised in isolation. The ASR module (Whisper-tiny) emitted 1–2 s transcript chunks with a median chunking delay of ~180–250 ms per window under the same conditions. Prosody feature extraction added ~8–15 ms per second of audio, and the lightweight fusion/overlay logic contributed <10 ms per event.

When these modules were run together in a best-effort integration harness on the target CPU-only configuration (720p), the system achieved near real-time behavior on short clips, but throughput and synchronization stability were variable under higher load. Approximate end-to-end alerting, from detected onset to on-screen notification, fell in the ~340–450 ms range on average, but occasional spikes were observed during ASR bursts and frame-rate dips, reflecting the absence of GPU acceleration and containerized runtime guarantees. Dropped-frame rates typically remained within ~2–3% in steady state yet increased during transitional resynchronization events.

Overall, the measurements indicate that the individual modules meet the latency targets in isolation, while the combined pipeline on CPU-only hardware remains sensitive to load and dependency interactions. These observations are consistent with the limitations outlined in Section 6.3 below and motivate the future engineering steps already identified:

GPU-enabled builds, containerization, and more robust stream synchronization before field deployment.

5.3 Detection Quality

Preliminary observations indicate that fusing visual, prosodic, and textual cues reduces false positives compared with any single modality, particularly around brief gaze aversions that coincide with positive verbal signals. Final analyses will compute precision, recall, F1, ROC-AUC, and time-alignment error for each event type, with 95% confidence intervals. Inter-rater reliability for ground truth will be reported (Cohen's κ), and disagreements adjudicated before scoring.

5.4 Usability and Interpretability

A short formative study will administer the SUS and a two-item perceived usefulness and explanation clarity questionnaire after participants conduct or observe a mock call using the dashboard. Qualitative comments will be coded thematically to refine message phrasing. The goal is not summative UX benchmarking but evidence that the prompts are legible, timely, and non-intrusive.

5.5 Robustness Checks

Ablation analyses will compare text-only, vision-only, and audio-only configurations to the fused system to quantify the contribution of each stream. Sensitivity analyses will test alternate sentiment analyzers and different smoothing windows to check that results are not artifacts of a single pre-processing choice.

5.6 Summary of Findings

Given the current state of the build, results are based on preliminary, component-level evaluations rather than full system integration. Early runs demonstrate that pose detection, sentiment analysis, and data export operate reliably in isolation, and that fusion logic improves event detection stability in controlled conditions. However, integrated end-to-end testing was not fully realized, and some findings remain indicative rather than conclusive. Further work will be required to validate full operational performance.

6 Discussion and Conclusion

6.1 Interpretation

The artefact demonstrates that practical, real-time behavioral analytics for virtual client meetings is achievable without cloud resources or opaque, heavy models. By privileging interpretability and privacy alongside accuracy, the system addresses key barriers to adoption in sales and relationship-management contexts. The engagement index and

rationale strings translate raw signals into domain-relevant prompts that can plausibly change conversational behavior in the moment.

6.2 Implications

For practitioners, the framework offers a path to augment coaching and call reviews with in-call guidance and structured post-call summaries. For researchers, it provides a modular testbed for studying multimodal fusion, explanation design, and the human–AI interface in high-stakes dialogue. The architecture invites substitutions of stronger models as they become available while preserving transparency.

6.3 Limitations and Threats to Validity

This study's primary limitation is that the artefact did not achieve sustained, production-grade end-to-end functionality across full-length live sessions within the timeframe. While individual modules for pose detection, sentiment analysis, and data export were validated in isolation, integration challenges prevented full real-time multimodal analysis during live testing. As a result, findings are based on component-level evaluations rather than complete system performance.

Technical constraints further limit generalizability. Real-time inference across multiple modalities on CPU-only hardware introduced latency and synchronization issues, and cross-platform dependency mismatches reduced reproducibility. The absence of GPU acceleration or a containerized deployment environment also restricted optimization options.

From a methodological standpoint, test data comprised the author's individual video interaction and publicly available meeting recordings, which, while diverse in conversational content, may not capture the full variability of real-world client interactions. Annotation of engagement events, despite inter-rater reliability checks, remains inherently subjective. The reliance on rule-based fusion, although interpretable, may underfit the nuanced behaviors and linguistic subtleties found in professional meetings. These factors together mean that the present evaluation offers a proof-of-concept demonstration of the framework's feasibility rather than a definitive measure of its effectiveness in production conditions. Future work should address these technical and ecological limitations before the system can be considered deployment-ready.

6.4 Future Work

Future development will focus on achieving stable, end-to-end real-time operation of the artefact. This will include re-engineering the media pipeline to unify audio and video capture within a single framework (e.g., relying solely on streamlit-webrtc for both streams), thereby eliminating threading conflicts and improving synchronization. Models for sentiment, emotion, and sarcasm detection will be optimized or quantized to reduce CPU load, and GPU acceleration will be incorporated where available to minimize latency.

To move this from a demo to a successful pilot, four practical steps are recommended: (1) package the app so it runs consistently with no setup surprises; (2) build in privacy by

default (i.e., clear consent, data-minimization, secure storage) to meet GDPR now and the EU AI Act as it lands; (3) add simple, privacy-safe checks and logs so issues are easy to spot and fix without exposing sensitive data; and (4) run short UI tests with sales and relationship managers to fine-tune the on-screen wording and thresholds. This keeps the tool understandable in real time and readies it for a small enterprise pilot.

To improve reproducibility and deployment readiness, the system environment will be containerized using Docker, ensuring consistent dependency resolution across platforms. The rule-based fusion module will be extended or replaced with a learned weighting mechanism trained on a larger, domain-specific annotated dataset. This will allow the system to capture more nuanced behavioral and linguistic cues without sacrificing interpretability.

Evaluation will expand beyond publicly available recordings to include real-world sales and client meetings, with appropriate consent and anonymization. This will improve ecological validity and allow for the measurement of business-relevant outcomes such as client satisfaction, engagement duration, and conversion rates. Cross-cultural and multilingual testing will also be prioritized to ensure broader applicability.

Finally, integration with common conferencing platforms via plug-ins or APIs will be explored, enabling seamless operation within existing workflows. This will be accompanied by iterative user testing to refine interface prompts and ensure the system's guidance is both actionable and non-intrusive in high-stakes communication scenarios. Further studies should also consider a model in which the focus is on proactive outcomes versus reactive guidance. The more the data can be adapted to the capabilities of future models, such as incorporating more robust body language, facial expressions, etc, the more effective it will be at influencing behavior and sentiment.

7 Conclusion

This work delivers a functional prototype of a multimodal, real-time analysis framework for virtual client interactions. It balances speed, interpretability, and privacy, and proposes an evaluation plan aligned to practical deployment. With further data and integration effort, the approach could mature into an assistive layer for sales and relationship management that is both effective and ethically grounded.

References

Casenave, E., & Schmitt, L. (2025). Comparing AI coaching and sales manager coaching: A construal-level approach. *Journal of Business Research*, *190*, 115241. https://doi.org/10.1016/j.jbusres.2025.115241

Curran, D. (2023). Surveillance capitalism and systemic digital risk: The imperative to collect and connect and the risks of interconnectedness. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231177621

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding Explainability: Towards Social Transparency in AI systems. *Proceedings of the 2021 CHI*

Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3411764.3445188

Essahraui, S., Lamaakal, I., Maleh, Y., El Makkaoui, K., Bouami, M. F., Ouahbi, I., Abd, A. A., Almousa, M., & Rodrigues, J. (2025). Human Behavior Analysis: A Comprehensive Survey on Techniques, Applications, Challenges, and Future Directions. *IEEE Access*, *13*, 128379–128419. https://doi.org/10.1109/access.2025.3589938

Frumosu, F. D., Lønfeldt, N. N., Mora-Jensen, A. .-R. C., Das, S., Lund, N. L., Pagsberg, A. K., & Clemmensen, L. K. H. (2022). Interpretability by design using computer vision for behavioral sensing in child and adolescent psychiatry. In *Proceedings of Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning*John, V., & Yasutomo Kawanishi. (2022). Audio and Video-based Emotion Recognition using Multimodal Transformers. *2022 26th International Conference on Pattern Recognition (ICPR)*. https://doi.org/10.1109/icpr56361.2022.9956730

Kordi, M. (2024). Surveillance Capitalism: The Transformation of Raw Online Data into Valuable Assets by High-Tech Companies—Is AI Governance a Threat or a Solution to Our Privacy Concerns? *Springer EBooks*, 401–416. https://doi.org/10.1007/978-3-031-58795-5_18

Li, R., Zhao, J., Hu, J., Guo, S., & Jin, Q. (2020). *Multi-modal Fusion for Video Sentiment Analysis*. https://doi.org/10.1145/3423327.3423671

Li, W.-C., Yang, C.-J., Liu, B.-T., & Fang, W.-C. (2021). A Real-Time Affective Computing Platform Integrated with AI System-on-Chip Design and Multimodal Signal Processing System. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. https://doi.org/10.1109/embc46164.2021.9630979

Muhammad, G., & Hossain, M. S. (2021). Emotion Recognition for Cognitive Edge Computing Using Deep Learning. *IEEE Internet of Things Journal*, *8*(23), 16894–16901. https://doi.org/10.1109/jiot.2021.3058587

Oluwafemi, I. O., Clement, T., Adanigbo, O. S., Gbenle, T. P., & Adekunle, B. I. (2021). A Review of Ethical Considerations in AI-Driven Marketing Analytics: Privacy, Transparency, and Consumer Trust. *International Journal of Multidisciplinary Research and Growth Evaluation*, *2*(2), 428–435. https://doi.org/10.54660/.ijmrge.2021.2.2.428-435

Patel, K. H. (2025). Ethical AI Integration in Salesforce: A Framework for Privacy, Fairness, and Accountable Implementation. *Journal of Computer Science and Technology Studies*, *7*(4), 580–585. https://doi.org/10.32996/jcsts.2025.7.4.68

Pathak, G., Pandey, D., Sonkar, N., & Kohli, P. (2025). *SEEING BEYOND WORDS: LEVERAGING COMPUTER VISION FOR INTERVIEWEE ANALYSIS IN AI-DRIVEN VIDEO INTERVIEWS*. https://doi.org/10.2139/ssrn.5250720

Rai, A. (2020). Explainable AI: from Black Box to Glass Box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rouast, P. V., Adam, M., & Chiong, R. (2019). Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE Transactions on Affective Computing*, *14*(8), 1. https://doi.org/10.1109/taffc.2018.2890471

Shareef, F., Ajith, R., Kaushal, P., & Sengupta, K. (2024). RetailGPT: A Fine-Tuned LLM Architecture for Customer Experience and Sales Optimization. *2nd International*

Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), 1390–1394. https://doi.org/10.1109/icssas64001.2024.10760685

Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., & Nanayakkara, S. (2020). Multimodal Emotion Recognition with Transformer-Based Self Supervised Feature Fusion. *IEEE Access*, *8*, 1. https://doi.org/10.1109/access.2020.3026823

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. https://doi.org/10.1109/tpami.2022.3183112

Wei, Y., Zhong, Z., & Gu, J. (2022). Human emotion based real-time memory and computation management on resource-limited edge devices. *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 487–492. https://doi.org/10.1145/3489517.3530490

Zhang, W., Qiu, F., Wang, S., Zeng, H., Zhang, Z., An, R., Ma, B., & Ding, Y. (2022). *Transformer-based Multimodal Information Fusion for Facial Expression Analysis*. ArXiv.org. https://arxiv.org/abs/2203.12367?utm_source=chatgpt.com