Signals of Aggression: Modelling Multimodal Cues and Perceptual Effects in Virtual Agents

Shaun Jing Heng Ong
Department of Electrical and Computer Engineering
National University of Singapore
Singapore, Singapore
e1300188@u.nus.edu

Aiden Tat Yang Koh
Department of Electrical and Computer Engineering
National University of Singapore
Singapore, Singapore
aiden@nus.edu.sg

Shaoyu Cai*
College of Design and Engineering
National University of Singapore
Singapore, Singapore
shaoyucai@nus.edu.sg

Felicia Fang-Yi Tan
Tandon School of Engineering
New York University
New York, New York, USA
felicia.tan@nyu.edu

Patrick Chia
Division of Industrial Design, School of Design and Environment
National University of Singapore
Singapore, Singapore
didcsl@nus.edu.sg

Eng Tat Khoo*
College of Design and Engineering
National University of Singapore
Singapore, Singapore
etkhoo@nus.edu.sg

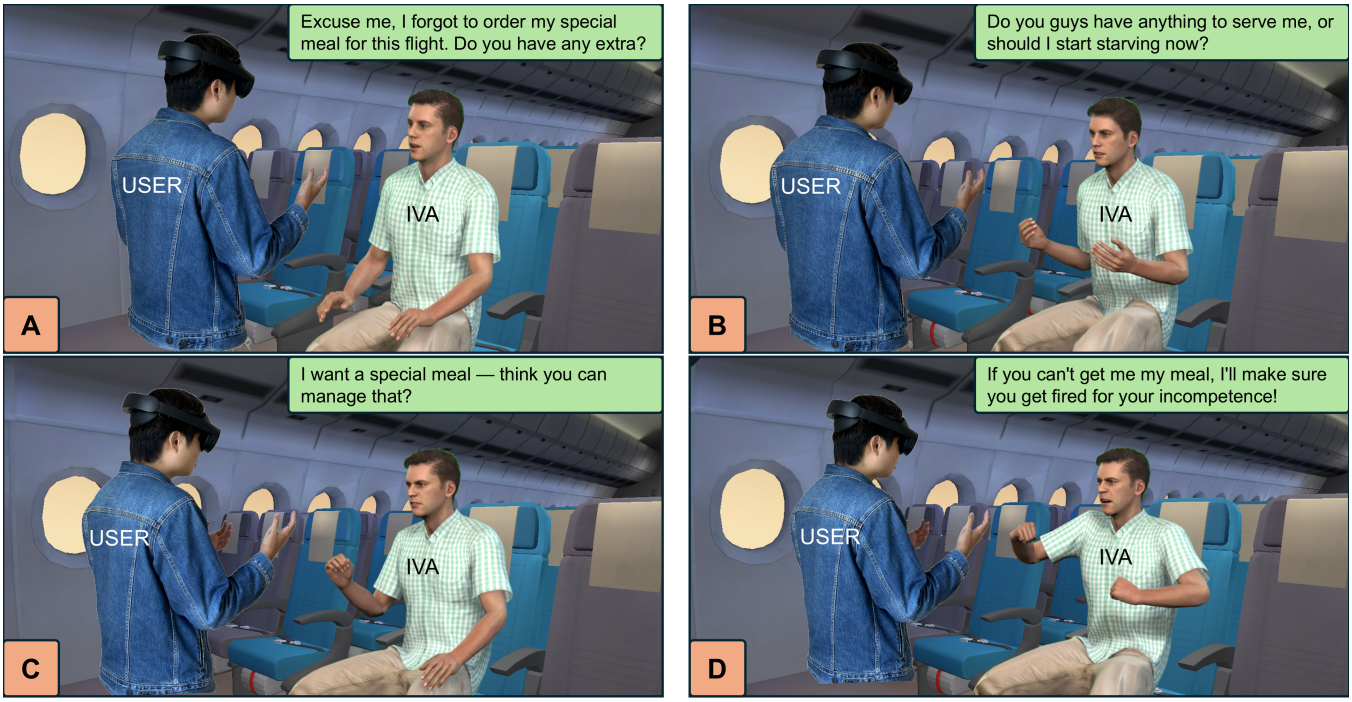


Figure 1: We present a multimodal aggression model to model aggression of an Intelligent Virtual Agent (IVA) by managing verbally aggressive message types used, roughness and loudness of the IVA's voice, anger-related facial action units and Laban Movement Analysis factors. The IVA expresses escalating levels of verbal and nonverbal aggression: (A) Not Aggressive, (B) Slightly Aggressive, (C) Aggressive, (D) Very Aggressive.

*Corresponding authors

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791031

Abstract

Aggression is a socially complex behaviour that intelligent virtual agents (IVAs) must convincingly convey in applications such as customer service and conflict training. Despite its importance, aggression remains understudied: prior work has focused on basic emotions and unimodal cues, providing little insight into how aggression can be modelled multimodally or systematically scaled

by intensity. We present a psychologically grounded model that parametrises language, voice, body movement and facial expressions, across four aggression levels. We evaluated the model in two studies with 38 flight attendants. Experiment 1 tested unimodal cues, showing all modalities except language conveyed aggression gradients. Experiment 2 extended this by combining modalities, demonstrating that coordinated multimodal integration stabilised weaker language cues and produced perceptually robust aggression levels (low, mid, and high) with body and facial cues carrying most weight. Our work contributes the first validated multimodal, multi-level aggression model for IVAs, offering design principles for broader socially expressive agents.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; Virtual reality; User studies; • **Computing methodologies** → Perception.

Keywords

Intelligent Virtual Agents, Virtual Reality, Multimodal Interaction, Aggression Modelling, Affective Computing, Language, Speech Prosody, Body Movement, Facial Expressions

ACM Reference Format:

Shaun Jing Heng Ong, Aiden Tat Yang Koh, Shaoyu Cai, Felicia Fang-Yi Tan, Patrick Chia, and Eng Tat Khoo. 2026. Signals of Aggression: Modelling Multimodal Cues and Perceptual Effects in Virtual Agents. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3772318.3791031

1 Introduction

Aggression in service professions is a growing concern across domains, from aviation to healthcare, where front-line staff frequently face verbal and non-verbal hostility from customers or clients [74]. Such incidents can harm employee well-being and undermine service quality [10, 13]. Addressing this problem requires training approaches that both prepare employees for the emotional impact of aggression and build effective de-escalation skills. Traditional approaches such as classroom-based instruction or peer role-play can be valuable, but they often lack the realism, repeatability, and range of scenarios needed to develop adaptive strategies [31].

Simulation-based training offers a way to create safe, controllable, and varied exposure to aggressive encounters. One promising approach uses Intelligent Virtual Agents (IVAs) within immersive virtual reality, leveraging advances in animation and speech synthesis to portray realistic emotional cues and behavioural dynamics. As IVAs are increasingly used in VR-based training, understanding how users perceive aggression in this context is essential for designing agents that feel convincing and elicit appropriate responses. For such training to be effective, IVAs must convincingly portray aggression—eliciting emotional reactions such as heightened alertness, stress, or recognition of threat [9].

While multimodal emotion perception has been widely studied in HCI and affective computing through wearable devices, expressive interfaces, and even inanimate objects [1, 38, 92, 94], comparatively few studies examine how humans perceive emotion in IVAs across multiple modalities—a gap especially relevant for training and simulation contexts. Prior work shows that language, voice, body movement, and facial expressions shape how virtual agents are perceived [7, 20, 61, 62], but little empirical evidence addresses how aggressive cues, and their varying intensities across modalities, influence perceptions of aggression.

Existing IVA-based training systems often prioritize scenario realism [11, 47, 60, 87], yet their aggression modelling remains unclear, with little attention to which modalities matter most, how cues should be scaled, or how they interact. This is important because prior work has shown that verbal and non-verbal cues can reinforce or weaken perceived affect depending on their intensity and alignment [18, 39, 44]. Without systematically examining how people perceive and integrate these cues, training systems risk leaving aggression signals ambiguous, making it harder for trainees to recognize or respond appropriately. Yet, to our knowledge, no study has comprehensively investigated the independent and combined effects of the three modalities—Language, Voice, and Body (combining body movement and facial expressions)—on perceived aggression in IVAs.

To address this gap, we examine how humans perceive aggression in an IVA for aviation customer service, whose behaviour is driven by a multimodal aggression model. Grounded in psychological characterizations of aggressive attributes (Sec. 3.1), the model varies verbally aggressive message types (Language), modulates vocal roughness and loudness (Voice), and incrementally adjusts anger-related facial expressions and body movements via Laban Movement Analysis (Body). These cues are synchronized in the IVA to generate realistic displays of aggression. We focus on everyday customer aggression that characterizes routine service interactions, rather than the more extreme or violent incidents typical of high-stakes security or emergency domains. Aviation provides a relevant and well-documented case: flight attendants regularly encounter escalating passenger frustration, verbal aggression, and non-verbal hostility, which constitute a significant occupational stressor with implications for safety, performance, and well-being [24]. Given the real-world relevance and intensity of these challenges, aviation provides an ideal context to investigate aggression perception in human–computer interaction. Therefore, we collaborated with a commercial airline and recruited 38 professional flight attendants as participants across two studies, to inform the design of IVA-based de-escalation training for service staff.

We address three research questions: RQ1 examines how effectively users perceive distinct levels of aggression from individual modalities; RQ2 explores how users integrate multiple cues—whether perceived aggression is additive, averaged, or otherwise influenced; and RQ3 investigates whether integrated multimodal cues can reliably convey distinct aggression levels. To answer these questions, we conducted two user-perception studies with 38 flight attendants. The first unimodality experiment examined how participants perceived aggression from Language, Voice, Body in isolation (RQ1), establishing baseline sensitivity to each modality and the clarity of aggression gradients. Participants reliably interpreted aggression across modalities, though higher levels of aggression were harder to distinguish for Language, leading us to collapse the top two Language levels for the subsequent study.

The second multimodality experiment assessed how participants perceived aggression when modalities were combined, to determine how users integrate signals (RQ2) and whether distinct aggression levels are maintained (RQ3). Participants consistently recognized three distinct aggression levels (*Low*, *Mid*, *High*) when cues were aligned, with body movement and facial expressions remaining the strongest predictors of perceived aggression. When cues conflicted, participants tended to average them, producing less distinct perceptions. These findings reveal how users interpret multimodal aggression and provide guidance for designing virtual agents that communicate aggression effectively in training contexts for front-line service professionals.

Our contributions are:

(1) A psychologically grounded multimodal aggression model for IVAs, integrating emotionally parametrised language, voice, body movement, and facial expressions. To our knowledge, this is the first model to systematically validate aggression cues across these three modalities.
(2) A two-stage experimental evaluation, comprising unimodal and multimodal user-perception studies, that examines how aggression cues in language, voice, body movement and facial expressions, are perceived individually and in combination.
(3) A set of design guidelines for creating aggressive IVAs that balance realism, clarity, and coherence across modalities, supporting the design of training simulations for front-line service staff, with principles that may extend to emotionally expressive agents used in other domains.

2 Related Work

Prior research on Intelligent Virtual Agents (IVAs) has focused heavily on multimodal emotion modelling, but aggression remains underexplored despite its importance for training in high-stakes contexts. We review three strands of work most relevant to our study: (1) emotion modelling in IVAs, (2) aggression as a target emotion, and (3) evaluation methods for aggression in IVAs.

2.1 Emotion Modelling in Intelligent Virtual Agents

IVAs are interactive, computer-generated characters used in domains ranging from training and education to customer service [14, 54]. Their effectiveness depends not only on producing context-appropriate verbal content but also on conveying emotions and behaviours in ways that feel believable to users. Emotion modelling in IVAs typically combines verbal and non-verbal cues — such as speech prosody, facial expressions, and body movement — to produce coherent emotional displays [32, 64, 67, 76, 84].

Another common theme is personality and context integration. Emotional displays are tailored by an agent's personality traits or social context to appear consistent and context-appropriate, providing holistic emotion models that combine multiple modalities [63, 75, 82, 84]. Randhavane et al. systematically modified parameters of gaits, gestures, and gaze to model a friendly IVA, showing how personality-driven behaviours can be embedded in movement [75]. Studies have also demonstrated that modifying an agent's gestures and tone according to a Big-Five personality profile yields more natural and engaging behaviour [84].

Overall, the trend is towards holistic emotion models that move beyond simplistic, single-channel approaches. This multimodal realism is seen as critical for believability, repeating early sentiments that truly lifelike agents must require coordinated verbal and non-verbal emotional cues [17, 39]. Existing work has emphasized modelling basic emotions or positive traits, but comparatively little attention has been given to how variations in emotional intensity are expressed and perceived.

2.2 Aggression as a Target Emotion

Among underexplored emotions, aggression stands out for its social complexity and practical importance. Prior work on IVA has examined aggression primarily in conflict or de-escalation training [9, 85, 87], but systematic emotion modelling remains scarce [30, 65].

A handful of studies have shown that aggressive agents can provoke stress and arousal comparable to live antagonists [9, 12], highlighting their potential for realistic training scenarios. The ability to represent aggression along a continuum is essential for IVAs to appear realistic and effective in such high-stakes interactions. Aggression is a high-arousal state often associated with harmful intent [2], expressed through distinctive behavioural cues. It may be reactive (impulsive, emotionally driven) or proactive (goal-oriented, instrumental) [52]. According to the frustration–aggression hypothesis [8, 25], frustration from blocked goals or unmet expectations can trigger anger and increase the likelihood of impulsive, emotionally charged aggression. This pathway aligns closely with reactive aggression, characterised by heightened arousal, reduced inhibitory control, and spontaneous, emotionally driven behaviours. Unlike proactive aggression, which is deliberate and planned, reactive aggression unfolds abruptly and is conveyed through rapid shifts in tone, posture, or phrasing.

In this work, we focus on routine, non-violent expressions of reactive aggression, such as irritated or accusatory language, raised vocal prosody, and tense, agitated body movements. Such behaviours are frequently observed in everyday interactions of customer service contexts, including aviation. These forms of aggression are thus prevalent and socially impactful, and modelling them accurately is essential for IVAs designed for conflict training, customer service simulation, or emotionally responsive interactions. Psychological models, such as Plutchik's Wheel of Emotions [72], highlight anger, a core component of aggression, as existing along a graded scale from mild irritation to intense rage. Despite this nuance, most IVA implementations treat aggression as a binary construct, often reducing it to a subset of anger or ignoring it altogether [21].

Modelling aggression on a graded scale is essential because its intensity strongly influences how others perceive threat and react behaviourally, or experience emotional impact. Even subtle differences in aggression can lead to very different social responses, making a binary representation inadequate. Capturing aggression as a graded, multimodal construct is therefore necessary to create behaviours that are both believable and impactful for the user.

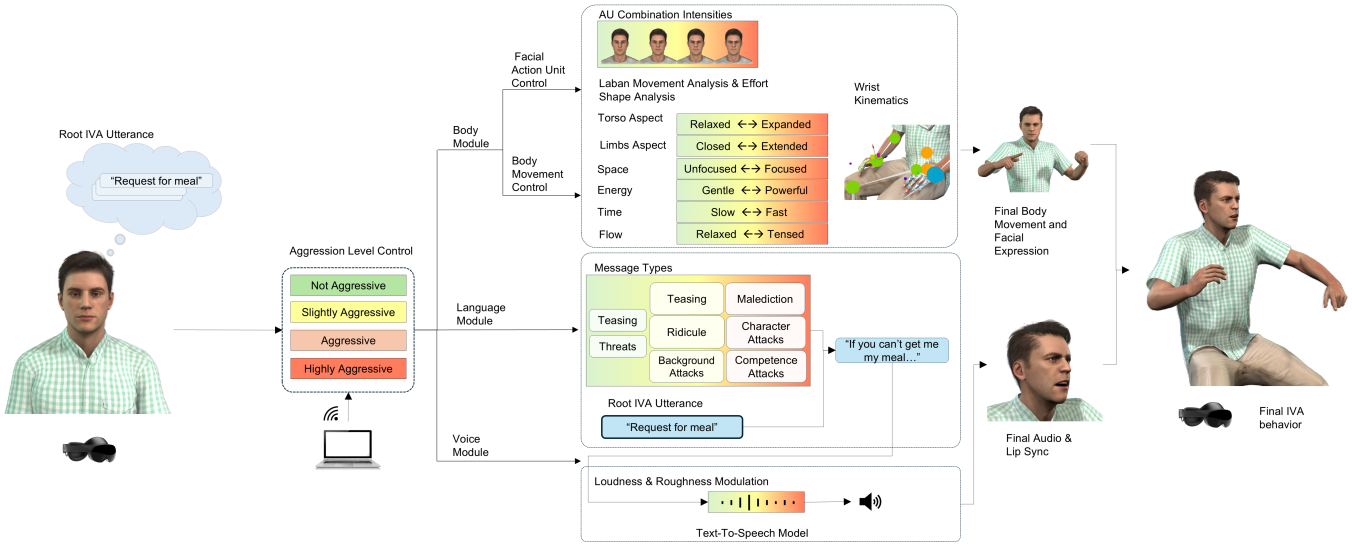


Figure 2: System framework of the IVA showing multimodal response generation: a neutral utterance is paraphrased to convey the target aggression level, converted to speech with lip-sync, and combined with corresponding facial and body animations to produce a cohesive expression of aggression.

2.3 Evaluating Aggression in IVAs

Designing IVAs that effectively convey graded aggression requires understanding (1) how humans display varying levels of aggression, and (2) how users would perceive these varying levels of aggression when they are expressed by an IVA. The first aspect informs what behaviours an aggressive IVA should enact at the varying levels of aggression, while the second examines whether users can reliably perceive these aggressive cues and infer the associated aggression level. In this work, prior research on how humans express aggression informs the behavioural parameters encoded in the IVA, and our user studies then examine how participants perceive these parametrised cues across modalities and levels.

Prior work characterises how humans express aggression across multiple behavioural channels: (1) Verbal cues ranging from mild teasing to direct personal attacks [3, 42, 43], (2) vocal cues such as increased loudness and spectral roughness, both of which are strong predictors of perceived anger [6, 69], (3) facial cues involving specific combinations of action units that have been empirically linked to anger expressions (e.g., AU4, AU5, AU7, AU23) [29, 50], and (4) body movement that conveys force, expansiveness, and directed movement, with "hot anger" often associated with high movement activity, spatial extension, and tension [22, 89].

While prior research describes how humans produce aggressive behaviours, it does not reveal how reliably users perceive these behaviours as aggressive when displayed by an IVA, especially in multimodal contexts. Prior work shows that multimodal combinations influence emotion perception [23, 78], but few studies have systematically examined how aggression levels can be parametrically varied within each modality, or how modalities integrate – additively, synergistically, or competitively – to shape perceived aggression in IVAs. These gaps motivate our investigation of modality-specific and multimodal aggression cues in a controlled IVA framework.

Most aggression evaluations to date have relied on indirect measures of physiological responses such as electrodermal activity, heart rate [9, 70], assuming stronger responses indicate more believable aggression. These physiological indicators provide objective evidence of arousal and stress but reveal little about how users consciously interpret aggressive cues. For training applications such as de-escalation, however, perception accuracy is critical: trainees must correctly recognize aggression levels to respond effectively.

In contrast, our work, inspired by Rodrigues et al. [78], directly assesses how participants perceive aggression levels exhibited by the IVA, addressing a critical gap in previous research. By focusing on users' conscious interpretations rather than physiological responses, we can examine how aggressive attributes across different modalities shape the perception of aggression.

3 Design of Aggression Model

This section outlines the design and implementation of our multimodal IVA system, which is capable of exhibiting varying levels of aggression through three modalities—Language, Voice, and a combined body movement & facial expression modality (referred to as Body).

3.1 Aggression Framework

3.1.1 Levels of Aggression. To model aggression in IVAs on a graded scale, we conceptualized it as varying in intensity, from annoyance and anger to rage, consistent with Plutchik's anger intensity spectrum [72]. Based on this framework, we defined four levels of aggression: *Not Aggressive*, *Slightly Aggressive*, *Aggressive*, and *Very Aggressive*. These levels guided the adjustment of aggressive cues

Table 1: Levels of Aggression and Their Message Types

Aggression Level	Message Types	Example Messages
Not Aggressive	No aggressive message types present	"I haven't received my order yet."
Slightly Aggressive	Teasing, threats	"Still waiting on that order—did it get lost on the way?"
Aggressive	Swearing, ridicule, background attacks	"This is ridiculous! Do you people ever get anything right?"
Very Aggressive	Maledictions, character attacks, competence attacks, physical appearance attacks	"Whoever handled my order should be ashamed of themselves! People like you are a disaster and should not be in this job. I hope you get fired for your incompetence."

in each modality—Language, Voice, and Body—so that the different levels of aggression were clearly distinguishable.

3.1.2 Language Aggression. We use Plutchik's spectrum [72] to define graded levels of aggression across modalities. For Language, we further draw on Infante's taxonomy to map concrete message types onto these four aggression levels. Designed to inflict psychological harm, verbally aggressive messages vary in severity—from mild teasing to severe competence attacks—each producing a different level of perceived hurt [40–43]. Prior work on chatbots and emotionally responsive IVAs has often relied on large language models (LLMs) conditioned on emotion labels to generate affective utterances [36]. However, such label-based approaches do not fully capture the nuanced, context-dependent nature of verbal aggression, especially across graded levels. By grounding our design in established communication taxonomies, we provide more precise and interpretable mappings between linguistic content and perceived aggression.

Following Infante et al.'s taxonomy of verbally aggressive messages [41], we classified our aggressive language into specific message types, such as teasing, threats, and character attacks, each mapped to a distinct aggression level. Table 1 summarizes the four levels of aggressive language. These categorizations were later validated in Experiment 1 (Unimodal Perception) to ensure they aligned with participants' perceptions.

To ensure that the language accurately reflects the intended level of aggression while maintaining its semantic context, the IVA's responses were generated using a chain-of-thought prompting approach [91] with a large language model (LLM). For each specified aggression level, the LLM (OpenAI's GPT-4) first produced a neutral utterance and then rephrased it to embed the corresponding aggressive message types.

The process involves several steps:

(1) **Starting Phrase Generation:** The system first generates a starting phrase relevant to the interaction context. For example, the virtual customer might say *"I haven't received my order yet."*
(2) **Aggression Level Determination:** The current aggression level informs the selection of message types that correspond to that level. For moderate aggression, message types like teasing and threats are selected.
(3) **Incorporation of Message Types:** Definitions and examples of the selected message types are incorporated into the prompt to guide the generation of the aggressive phrase.
(4) **Aggressive Phrase Generation:** The starting phrase (Neutral) is paraphrased and reformulated so that it carries the linguistic markers of the target aggression level (e.g., adding teasing, threats, or attacks). It may result in a statement like, *"Still waiting on that order—did it get lost on the way?"*

3.1.3 Voice Aggression. Research has shown that variations in acoustic features play a crucial role in how listeners perceive emotions, particularly anger. Pell et al. [69] and Banziger et al. [6] identified high acoustic intensity (perceived as loudness) and shifts in spectral balance (perceived as roughness) as key indicators of anger. Based on these findings, we designed vocal expressions for the IVA with systematically varied loudness and roughness to convey different levels of aggression. We synthesized neutral pseudo-utterances using Azure Text-to-Speech (TTS) with Custom Neural Voice models, based on Uni-TTS architecture [59] and style transfer [58]. By adjusting the *StyleDegree* parameter (0.0 = *Not Aggressive* to 2.0 = *Very Aggressive*), we modulated perceived aggression in the synthesized voices. Table 2 shows the correspondence between *StyleDegree* values and aggression levels. These vocal manipulations were later evaluated in Experiment 1 (Unimodal Perception) to confirm that participants' perceptions of the vocal cues aligned with the intended aggression levels.

Table 2: Mapping of Aggression Levels to Azure TTS *StyleDegree* control parameters

Aggression Level	StyleDegree
Not Aggressive	0
Slightly Aggressive	0.67
Aggressive	1.33
Very Aggressive	2

3.1.4 Body Aggression. Previous works investigated emotion perception based on facial and bodily expressions. For example, Meeren et al. [55] demonstrated that observers rapidly and automatically integrate facial and bodily expressions during emotion perception, highlighting the importance of combining both modalities. Similarly, Aviezer et al. [4] found that body cues significantly influence the perception of facial expressions, especially when facial expressions are ambiguous or subtle. Building upon this understanding, we designed our IVA to integrate both facial expressions and body animations, as both are significant contributors to emotion perception through physical cues.

Facial Expressions. Several observer-based systems for measuring facial expressions have been developed. [27, 28, 45]. Among these various systems, the Facial Action Coding System (FACS) stands out as the most comprehensive, methodologically rigorous, and widely used [28]. The FACS allows the coding of nearly all observable facial expressions, which are decomposed into action units (AUs), which represent the smallest visually discernible facial movements.

FACS is inherently descriptive: it codes facial muscle movements (AUs) without assigning them to specific emotions. Emotion-specific interpretations emerge from empirically validated combination rules, as outlined in the FACS Manual and Investigators' Guide [29]. In the context of anger, a key component of aggression, the most commonly recognized AU combinations are AU4, AU5, AU7, and AU23. These can appear in different combinations and intensities, with strong expressions featuring intense AU4, AU5, and AU7, and milder expressions involving less intense AUs [50, 51].

To generate standardized AU combinations across intensities for our IVA, we utilized FACSGen, a validated tool that synthesizes facial expressions through the systematic manipulation of AU units [79]. Facial expressions of the required intensities are seen in Fig. 3.

Figure 3: Facial expressions of our IVA depicting increasing intensities of action units (AUs) from 0% (a), 33.3% (b), 66.6% (c), to 100% (d). These increments illustrate the progressive activation of AUs, with (a) showing *Not Aggressive* expression and (d) representing *Very Aggressive*.

Body Movement.

To guide our generation of standardized aggressive body movements, we applied Laban Movement Analysis (LMA), Effort-Shape Analysis, and wrist kinematics.

LMA is a comprehensive framework used for describing and analysing human movement, providing organised components and nuanced descriptions of the scope of human movement. LMA has four main components; Body, Effort, Shape and Space. Body and Space describe the structural aspects of movement, while Effort and Shape address its qualitative and expressive characteristics. Effort specifies the deployment of energy along four bipolar dimensions; Weight, Time, Space and Flow, thereby capturing the dynamic modulation of movement. Shape characterizes the evolving configuration of the body in motion, comprising Shape Flow, Directional change, and Carving/Shaping actions, thus providing a comprehensive account of human movement [35, 83]. Melzer et al. has shown that anger is associated with specific Laban components, describing anger as an approach (advancing) emotion and is characterized by Strong weight, Sudden timing, Advancing direction, and Direct focus [57]. The combination of these three specific efforts is also known in LMA as an action drive, and characterizes purposeful movements and actions that are driven by a certain aim [81].

Wallbott et al. further expanded on this action drive, distinguishing "cold anger" (low expansiveness, low energy) from "hot anger" (expansive, high energy). "Hot anger," or full-blown anger, often involves lifting of the shoulders, lateralized hand/arm movements, and opening and closing of the hands. Such movements are typically characterized by high movement activity, expansiveness, spatial extension, and energy. In contrast, "cold anger," a subtler form of anger, exhibits lower values for these movement patterns [89].

Table 3: Body Movement Across Aggression Levels

Dimension	Not Aggressive	Slightly Aggressive	Aggressive	Very Aggressive
Torso [22, 35, 83]	Relaxed, neutral	Slightly expanded	Expanded, stretched	Highly expanded
Limb [22, 35, 83, 89]	Close to body	Slightly away, minimal movement	Expanded, moderate movement	Fully extended, vigorous
Space [57, 81]	Indirect, unfocused	Somewhat direct	Direct, focused	Very direct, highly focused
Energy [22, 57, 89]	Light, gentle	Moderate energy	Strong, forceful	Very strong, powerful
Time [57, 81]	Sustained, slow	Somewhat hurried	Sudden, fast	Very sudden, very fast
Flow [22, 35, 83]	Free, relaxed	Slightly bound	Bound, tense	Highly bound, very tense
Movement Characteristics [22, 89]	Minimal hand motion	Slight shoulder lift, subtle hand/arm movements	Noticeable shoulder lift, moderate hand/arm activity	High shoulder lift, vigorous hand/arm activity
Avg.Wrist Velocity (m/s) [73]	0.30	0.425	0.55	0.70
Peak Wrist Accel. (m/s^2) [73]	1.0	1.5	2.0	2.5
Peak Wrist Jerk (m/s^3) [73]	5.0	7.5	10.0	12.5

Effort-Shape Analysis is often used to describe the movement style characteristics associated with target emotions, including anger. The movement style characteristics for anger are derived from Crane et al. [22], who specifically analyse the movement qualities associated with emotions like anger by applying LMA principles, focusing on Effort factors such as Weight, Time, Space, and Flow, as well as Shape aspects like torso and limb movements.

Pollick et al. [73] demonstrated that tangential velocity, acceleration, and jerk of wrist movements are linked to perceived emotions. By incrementally increasing these kinematic properties, we created a quantifiable gradient of aggression intensity across the body animations. These adjustments reflect the Effort-Time (suddenness) and Effort-Weight (strength) factors from LMA, correlating with more aggressive emotional expressions.

We combined Effort-Shape descriptors, movement characteristics, and wrist kinematic properties to create the four different aggression levels as shown in Table 3. Based on this, we created the IVA's body movement animations using motion capture techniques using video-to-animation technology [71], and adjusted the kinematics of the motion capture data with 3D animation software [77].

The varying levels of Body aggression intensities are illustrated in Fig. 4, which we later validated against participants' perceptions in Experiment 1 (Unimodal Perception).

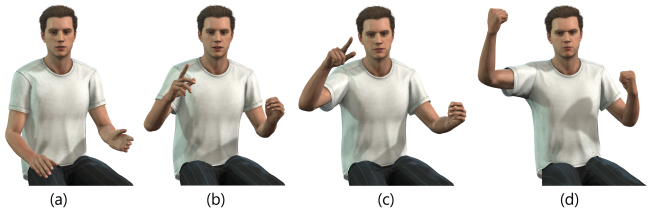


Figure 4: Body Movement and Facial Expressions of an IVA depicting increasing intensities across the four aggression levels: (a) *Not Aggressive*, (b) *Slightly Aggressive*, (c) *Aggressive*, (d) *Very Aggressive*.

3.2 Model Implementation

To integrate responses across modalities in VR, the IVA's verbal output was first generated as a neutral utterance by the LLM (OpenAI's GPT-4), paraphrased to convey the intended message types, and converted to speech via Azure Text-to-Speech using the *StyleDegree* parameter corresponding to the aggression level, with audio synchronized to mouth movements using lip-sync. Facial expressions and body movements for different aggression levels were implemented using pre-designed animations in Unity3D, with the animation controller enabling selection and blending based on the IVA's emotional state. Audio and visual components were synchronized to ensure a cohesive multimodal experience. The system framework, illustrating how varying aggression levels were conveyed, is shown in Fig. 2.

Importantly, because the IVA's verbal, vocal, and bodily cues can be modified in real time via a networked client, its aggression level can be dynamically adjusted throughout an interaction. This design makes the system readily extensible: the current manual control mechanism can be replaced by an adaptive layer that maps user responses or contextual signals, such as per-turn performance or indicators of user stress, to real-time adjustments in aggression, providing a flexible foundation for the development of interactive, real-time adaptive IVA systems.

4 Experiment 1 (Unimodal Perception)

To design effective virtual agents for customer service training, it is essential to validate how users perceive aggression. In the unimodal experiment (Experiment 1), we recruited flight attendants, whose professional experience with interpersonal interactions provides a relevant context for studying aggression perception. We examined how participants perceived the four levels of aggression when conveyed through Language, Voice, and Body cues. This experiment also validated the modality-specific aggression parameters in Sec. 3.1 and laid the groundwork for studying combined cues in multimodal aggression modelling in Experiment 2 (Multimodal Perception). The research question guiding this study was: How effectively do users perceive distinct levels of aggression from each modality?

4.1 Participants

19 flight attendants (14 female, 5 male), aged 23 to 56 years (M = 28.8, SD = 8.7), took part in the study. Their professional experience ranged from 2 to 32 years (M = 6.1, SD = 7.3), representing the typical demographic and experience distribution of cabin crew available on commercial flights operated by our airline partner. Recruitment was conducted in collaboration with the airline as part of a joint research initiative with the university. Participants received compensation in accordance with the company's internal policies for training or research participation. All participants were fluent in English and routinely used the language in professional settings. The experimental protocols were reviewed and approved by the Institutional Review Board (IRB).

4.2 Experimental Design

The uni-modality experiment investigated how aggression is perceived in three separate modalities: Language, Voice, and Body (including Body Movement & Facial Expression). A within-subject design was used, with aggression level as the independent variable. Each participant was exposed to four conditions within each modality: *Not Aggressive*, *Slightly Aggressive*, *Aggressive*, and *Very Aggressive*. The order of modalities as well as aggression conditions within each modality was fully counterbalanced. Participants' subjective ratings of perceived aggression served as the dependent variable.

In line with RQ1, experiment 1's objective was to validate aggression scaling within each modality independently. To meet this goal, the study was not structured as a *Modality × Aggression Level* factorial design. Isolating modalities established clean perceptual baselines without cross-modal influence, which Experiment 2 then uses to examine cross-modal integration.

4.3 Apparatus and Materials

To evaluate aggression perception in each modality independently, we created distinct stimuli for Language, Voice, and Body and presented them in isolation using a VR headset (Meta Quest Pro). This approach allowed us to measure participants' responses without cross-modal influence.

4.3.1 Language Modality. Four phrases representing the four aggression levels were generated using the process described in Sec. 3.1.2 (see Appendix A.1 for examples). Participants read each phrase in textual format within the VR environment, isolating aggression perception to linguistic content.

4.3.2 Voice Modality. Four synthetic audio files representing the four aggression levels were generated using Azure Custom Voice by linearly varying the *StyleDegree* parameter, as described in Sec. 3.1.3. Following prior work [5, 68], emotionally inflected pseudo-utterances were presented to participants to ensure that perceived changes in aggression were driven by vocal prosody rather than semantic content. Participants listened to each audio clip via headphones in the VR headset.

4.3.3 Body Modality. Four body animations and facial expressions representing the four aggression levels were created using specific combinations of facial AU intensities and body movement parameters, as described in Sec. 3.1.4. Participants viewed each animation within the VR environment, with full access to body and facial cues.

4.4 Procedure

After providing informed consent, participants completed the unimodality experiment for all three modalities in their assigned counterbalanced order, following a structured procedure:

(1) Participants were briefed on the experiment's purpose and given modality-specific instructions.
(2) Two example stimuli were presented to familiarize participants with the task and reduce initial bias or confusion.
(3) Participants were then presented with stimuli representing the four aggression conditions in their assigned counterbalanced order (see Sec. 4.3 for details).
(4) After each stimulus, participants completed a questionnaire rating perceived aggression and modality-specific aggressive attributes.

In total, participants completed 18 trials (6 stimuli × 3 modalities). The experiment lasted approximately 30 minutes per participant, with a 5-minute break between modalities to reduce fatigue from prolonged VR headset use.

4.5 Measures & Hypotheses

4.5.1 Perceived Aggression. Participants rated each stimulus on a 5-point Likert scale (1 = Least Aggressive, 5 = Most Aggressive) immediately after exposure. This measure captured participants' subjective ratings of perceived aggression for each modality and allowed us to evaluate whether the stimuli successfully represented the intended aggression levels.

H1: Perceived Aggression in Unimodal Conditions. We hypothesize that participants will perceive systematically increasing aggression across the four levels (*Not Aggressive* < *Slightly Aggressive* < *Aggressive* < *Very Aggressive*) for each modality (Language, Voice, Body). This hypothesis serves as a validation of our design rationale, establishing perceptual grounding for the multimodal integration tested in Experiment 2.

4.5.2 Modality-Specific Cues. In addition to overall aggression ratings, we collected ratings on modality-specific cues to validate that the attributes manipulated in each modality were perceived as intended. Each set of measures directly corresponded to the design parameters described in Sec. 3.1.

(1) **Language:** Participants selected all message types they perceived to be present in each verbal stimulus (e.g., teasing, threats, character attacks) using a multiple-response checklist derived from Infante's taxonomy [40–42]. A multi-response format was used because a single verbal phrase can convey multiple message types; a forced-choice design would impose artificial exclusivity, forcing a single "best-of-fit" selection and risking information loss and misrepresentation of participants' perceptions [80]. These ratings allowed us to assess whether participants detected the intended combination of message types encoded at different aggression levels.
(2) **Voice:** Participants rated how *loud* and *rough* each vocal stimulus sounded on two separate 5-point Likert scales (1 = Very Low, 5 = Very High), corresponding directly to the manipulated acoustic parameters (intensity and spectral balance). These ratings allowed us to assess whether participants perceived the intended acoustic manipulations.
(3) **Body:** Participants rated perceived movement qualities on multiple 5-point Likert scales (1 = Very Low, 5 = Very High), covering: (a) effort qualities (weight, time, space), (b) shape qualities (torso/limb expansion), and (c) facial expressions (e.g., eyebrow lowering, eyelid raising, eyelid tightening, lip tightening). These ratings allowed us to verify whether participants detected the intended FACS- and LMA-based manipulations in the animations.

4.6 Results

4.6.1 Perceived Aggression. The effects of the three modalities-Language, Voice, and Body—on perceived aggression of the virtual avatar were analysed independently. To assess the normality assumption for the one-way repeated measures ANOVA, the residuals for each modality were tested using the Shapiro-Wilk test (refer to Appendix A.3). These tests indicated significant deviations from normality across all modalities (all $p < 0.001$). Given this, and the ordinal nature of the Likert scale ratings, non-parametric Friedman tests were conducted to examine how different aggression levels within each modality affected perceived aggression. For any significant main effects, post hoc pairwise comparisons were performed using Wilcoxon signed-rank tests with Holm-Bonferroni adjustment.

Significant effects found across all three modalities. The Friedman test showed a significant effect of Language on perceived aggression ($\chi^2(3) = 51.12$, $p < 0.0001$), with Kendall's W = 0.90. Mean ratings increased across levels, with substantial rises from *Not Aggressive* ($M = 1.15$, $SD = 0.48$) to *Slightly Aggressive* ($M = 3.05$, $SD = 1.15$) and to *Aggressive* ($M = 4.79$, $SD = 0.52$), but only a slight increase to *Very Aggressive* ($M = 4.84$, $SD = 0.37$). Post hoc tests revealed significant differences in perceived aggression ratings between adjacent levels (*Not Aggressive-Slightly Aggressive*: $p < 0.01$; *Slightly Aggressive-Aggressive*: $p < 0.01$). However, no significant difference was found between *Aggressive* and *Very Aggressive*, indicating that participants' perception of aggression plateaued at higher intensities of Language cues. All other comparisons were significant ($p < 0.01$).

A significant effect of Voice on perceived aggression was found ($\chi^2(3) = 45.88$, $p < 0.0001$, Kendall's $W = 0.80$). Mean ratings increased steadily across levels (*Not Aggressive*: $M = 1.32$, $SD = 0.58$; *Slightly Aggressive*: $M = 2.68$, $SD = 1.11$; *Aggressive*: $M = 3.21$, $SD = 1.08$; *Very Aggressive*: $M = 4.53$, $SD = 0.84$). Post hoc tests confirmed that each successive level was perceived as significantly more aggressive than the previous one (*Not Aggressive–Slightly Aggressive*: $p < 0.001$; *Slightly Aggressive–Aggressive*: $p < 0.05$; *Aggressive–Very Aggressive*: $p < 0.05$), showing that participants reliably interpreted increasing aggression in Voice cues across all levels. All other comparisons were significant ($p < 0.01$).

Body also had a significant impact on perceived aggression ($\chi^2(3) = 49.72$, $p < 0.0001$, Kendall's $W = 0.87$). Mean ratings increased steadily across levels (*Not Aggressive*: $M = 1.11$, $SD = 0.46$; *Slightly Aggressive*: $M = 1.74$, $SD = 0.65$; *Aggressive*: $M = 3.21$, $SD = 1.08$; *Very Aggressive*: $M = 4.53$, $SD = 0.84$). Post hoc comparisons revealed significant differences between each adjacent level (*Not Aggressive–Slightly Aggressive*: $p < 0.01$; *Slightly Aggressive–Aggressive*: $p < 0.01$; *Aggressive–Very Aggressive*: $p <$

Table 4: Detailed comparisons of the three modalities based on participants' subjective ratings (mean, SD).

Language							
Perceived Aggression	NA (1.15, 0.48)	<	SA (3.05, 1.15)	<	A (4.79, 0.52)	∼	VA (4.84, 0.37)
Voice							
Perceived Aggression	NA (1.32, 0.58)	<	SA (2.68, 1.11)	<	A (3.47, 1.22)	<	VA (4.21, 0.92)
Loudness	NA (2.21, 0.97)	∼	SA (2.84, 1.17)	∼	A (3.53, 1.35)	<	VA (4.37, 1.07)
Roughness	NA (1.89, 1.05)	∼	SA (2.63, 1.01)	∼	A (3.26, 1.19)	∼	VA (3.84, 1.12)
Body Movement & Facial Expressions							
Perceived Aggression	NA (1.11, 0.46)	<	SA (1.74, 0.65)	<	A (3.21, 1.08)	<	VA (4.53, 0.84)
Effort-Weight	NA (1.11, 0.46)	<	SA (1.68, 0.75)	<	A (3.26, 0.99)	<	VA (4.37, 0.83)
Effort-Time	NA (1.37, 0.83)	<	SA (2.05, 0.62)	<	A (3.16, 1.26)	<	VA (4.42, 0.84)
Effort-Space	NA (1.58, 1.02)	<	SA (2.21, 0.79)	<	A (3.00, 0.94)	<	VA (4.53, 0.77)
Shape	NA (1.32, 0.67)	<	SA (2.05, 0.91)	<	A (3.32, 1.00)	<	VA (4.53, 0.70)
Lowering Eyebrows	NA (1.47, 0.90)	∼	SA (2.11, 1.29)	∼	A (2.84, 1.21)	<	VA (3.84, 1.12)
Raising Upper Eyelids	NA (1.68, 0.95)	∼	SA (2.16, 0.90)	∼	A (2.53, 1.26)	<	VA (3.74, 1.05)
Tightening Eyelids	NA (1.32, 0.58)	<	SA (2.11, 0.88)	∼	A (2.58, 1.02)	<	VA (3.74, 0.93)
Tightening Lips	NA (1.47, 1.02)	∼	SA (2.21, 1.27)	∼	A (3.05, 1.18)	∼	VA (3.63, 1.12)

NA = Not Aggressive, SA = Slightly Aggressive, A = Aggressive, VA = Very Aggressive.
'<' denotes significant difference ($p < 0.05$); '∼' indicates no significant difference.

0.001), indicating that participants distinguished Body cues clearly across aggression levels. All other comparisons were significant ($p < 0.001$).

Fig. 5 presents box plots illustrating the distribution of perceived aggression ratings across four aggression levels for each of the three modalities.

4.6.2 Validating aggressive attributes in each modality. For each modality, we analysed participants' ratings to confirm that modality-specific cues were perceived as intended and that the manipulated attributes contributed meaningfully to perceived aggression. Shapiro–Wilk tests indicated significant deviations from normality for all manipulated attributes (all $p < .010$), with the exception of lip tightening (refer to Appendix A.3). Accordingly, non-parametric Friedman tests were conducted for each attribute, followed by Wilcoxon signed-rank tests with Holm–Bonferroni correction for post hoc pairwise comparisons. Table 4 shows the detailed pairwise comparisons of each modality and its associated aggressive attributes.

Language cues. As responses were not mutually exclusive, statistics that rely on a single response per trial (e.g., forced-choice accuracy) cannot be meaningfully applied. Instead, we analysed message type recognition using endorsement rate (i.e. the percentage of participants selecting each category) which measures recognition frequency rather than categorical correctness and is the standard approach for multi-response survey items [16, 86]. Response patterns show that participants made selective judgments: 0 to 2 message types (M = 0.36, SD = 0.68) for *Not Aggressive* stimuli, 0 to 6 message types (M = 2.16, SD = 1.38) for *Slightly Aggressive*, 2 to 9 message types (M = 4.63, SD = 2.09) for *Aggressive*, and 2 to 9 message types (M = 4.63, SD = 2.24) for *Very Aggressive*, with only 2 participants ever selecting all options for a given stimulus, supporting the validity of the measure.

To assess how closely participants' selections matched the intended message types at each aggression level, we computed the Jaccard similarity [33] between their selections and the intended set of message types. Similarity was 0.74 ± 0.44 for *Not Aggressive* stimuli, 0.56 ± 0.31 for *Slightly Aggressive* stimuli, 0.26 ± 0.10 for *Aggressive* stimuli, and 0.48 ± 0.20 for *Very Aggressive* stimuli. The low similarity at the *Aggressive* level resulted from participants often selecting message types beyond those intended, such as "Character Attack" and "Competence Attack," which were meant for the *Very Aggressive* level. This pattern suggests a tendency to over-select or misidentify message types at higher aggression levels.

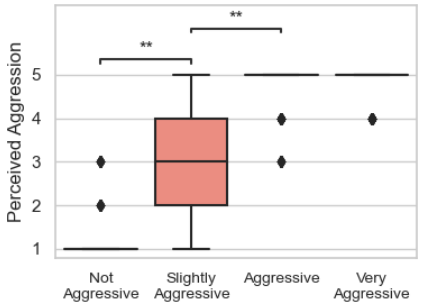
At the *Slightly Aggressive* level, "Teasing" and "Threats" were most frequently endorsed. At the *Aggressive* level, participants most often selected "Swearing", "Ridicule" and "Background Attack". At the *Very Aggressive* level, the most common selections were "Malediction", "Character Attack", "Competence Attack" and "Physical Appearance Attack". Overall, these findings indicate that the categorization of message types was largely perceived as intended and that participants could distinguish the manipulations across aggression levels, supporting our use of linguistic cues and their role in conveying graded aggression. Table 5 presents the heatmap of endorsement percentages along with the Jaccard similarity for each aggression level.

Voice cues. Friedman tests on participants' ratings of vocal cues, specifically loudness and roughness, revealed a significant effect of aggression level on both loudness ($\chi^2(3) = 32.75$, $p < 0.0001$, Kendall's $W = 0.57$) and roughness ($\chi^2(3) = 26.60$, $p < 0.0001$, Kendall's $W = 0.47$). Mean ratings increased with aggression levels for loudness (*Not Aggressive*: M = 2.21, SD = 0.97; *Slightly Aggressive*: M = 2.84, SD = 1.17; *Aggressive*: M = 3.53, SD = 1.35; *Very Aggressive*: M = 4.37, SD = 1.07) and roughness (*Not Aggressive*: M = 1.89, SD = 1.05; *Slightly Aggressive*: M = 2.63, SD = 1.01; *Aggressive*: M = 3.26, SD = 1.19; *Very Aggressive*: M = 3.84, SD = 1.12). For successive aggression levels, post hoc tests revealed a significant difference only between *Not Aggressive* and *Slightly Aggressive* for loudness. However, all other pairwise comparisons for loudness and roughness were significant ($p < 0.05$).

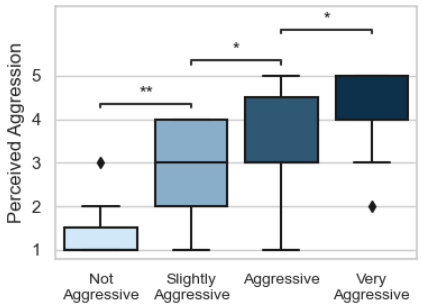
To assess whether vocal cues were meaningfully related to perceived aggression, we examined correlations between participants' perceived aggression and their ratings of loudness and roughness using Spearman's rank correlation. This analysis revealed strong

Table 5: Heat map showing the percentage of participants endorsing each message type across aggression levels, with Jaccard similarity reported in Mean (SD).

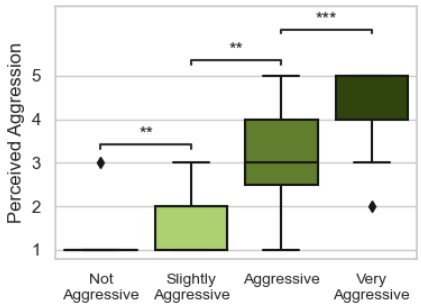
	Teasing	Threats	Swearing	Ridicule	Background Attack	Malediction	Character Attack	Competence Attack	Physical Appearance Attack	Jaccard Similarity Mean (SD)
Not Aggressive	15.79%	15.79%	0.00%	0.00%	0.00%	0.00%	0.00%	5.26%	0.00%	0.74 ± 0.44
Slightly Aggressive	73.68%	68.42%	21.05%	26.32%	0.00%	5.26%	5.26%	15.79%	0.00%	0.56 ± 0.31
Aggressive	31.58%	42.11%	78.95%	73.68%	84.21%	15.79%	47.37%	73.68%	15.79%	0.26 ± 0.10
Very Aggressive	21.05%	42.11%	36.84%	36.84%	42.11%	52.63%	68.42%	89.47%	73.68%	0.48 ± 0.20



(a) Language



(b) Voice



(c) Body Movement and Facial Expression

Figure 5: Perceived aggression ratings for all modalities across different levels of aggression

positive correlations for both loudness ($R = 0.79$, $p < 0.0001$) and roughness ($R = 0.83$, $p < 0.0001$), indicating that even though participants did not reliably distinguish all pairwise levels, their overall perception of aggression closely tracked these vocal attributes. These results support the effectiveness of our vocal manipulations in conveying graded aggression levels.

Body Cues. Friedman tests were conducted on participants' ratings of aggressive attributes related to Body. Significant effects were observed across all body movement attributes, including effort-weight ($\chi^2(3) = 50.80$, $p < 0.0001$, Kendall's $W = 0.88$), effort-time ($\chi^2(3) = 46.48$, $p < 0.0001$, Kendall's $W = 0.82$), effort-space ($\chi^2(3) = 44.33$, $p < 0.0001$, Kendall's $W = 0.78$), shape ($\chi^2(3) = 47.67$, $p < 0.0001$, Kendall's $W = 0.84$), as well as facial attributes such as lowering eyebrows ($\chi^2(3) = 28.09$, $p < 0.0001$, Kendall's $W = 0.49$), raising upper eyelids ($\chi^2(3) = 32.18$, $p < 0.0001$, Kendall's $W = 0.56$), tightening eyelids ($\chi^2(3) = 40.41$, $p < 0.0001$, Kendall's $W = 0.71$), and tightening lips ($\chi^2(3) = 32.88$, $p < 0.0001$, Kendall's $W = 0.58$). For successive aggression levels, post hoc tests revealed significant differences between aggression levels for all Body cues, except for some facial expressions. Specifically, no significant differences were found between *Not Aggressive* and *Slightly Aggressive*, or between *Slightly Aggressive* and *Aggressive* for lowering eyebrows and raising upper eyelids. Additionally, tightening eyelids did not differ significantly between *Slightly Aggressive* and *Aggressive*. No significant differences were also found for tightening lips. However, all other pairwise comparisons for the body movement and facial attributes were significant ($p < 0.05$).

To assess whether body cues were meaningfully associated with participants' perception of aggression, we examined correlations between perceived aggression and each body attribute using Spearman's rank correlation. Most body movement attributes showed strong positive correlations (effort-weight: $R = 0.94$, effort-time: $R = 0.85$, effort-space: $R = 0.83$, shape: $R = 0.88$, tightening eyelids: $R = 0.81$, all $p < 0.0001$), while facial-expression-related attributes exhibited moderate positive correlations (lowering eyebrows: $R = 0.70$, raising upper eyelids: $R = 0.70$, tightening lips: $R = 0.69$, all $p < 0.0001$). These findings suggest that participants were more sensitive to the manipulated body movement attributes than to facial cues, consistent with Aviezer et al. [4], who found that body cues dominate perception over more subtle facial cues. Despite somewhat weaker correlations for facial cues, both body and facial cues were strongly associated with perceived aggression, supporting the effectiveness of our manipulations in conveying graded aggression levels.

4.7 Discussion

H1: Perceived Aggression in Unimodal Conditions. We expected that participants would perceive a clear and distinguishable aggression gradient in each modality, with perceived aggression increasing systematically across levels: *Not Aggressive* < *Slightly*

Aggressive < *Aggressive* < *Very Aggressive*. Our findings largely support this hypothesis. Participants consistently perceived higher levels of aggression as the intensity of the cues increased across all three modalities, but the clarity of this gradient varied by modality. Body and Voice cues were perceived distinctly across all four aggression levels, whereas Language cues exhibited greater perceptual variability.

Specifically, participants were able to differentiate lower aggression levels in Language (*Not Aggressive*, *Slightly Aggressive*, and *Aggressive*), but they did not reliably distinguish between *Aggressive* and *Very Aggressive*, suggesting a plateau in perceived aggression at higher intensities. This plateau might have resulted from two factors. Firstly, the Jaccard similarity results indicated that participants often identified more message types than were actually present for the *Aggressive* level. Exposure to higher levels of aggressive language may have triggered a hostile attribution bias [19], causing participants to perceive message types that were not intended. This over-selection of message types by participants may have contributed to a similar level of perceived aggression between *Aggressive* and *Very Aggressive* stimuli. Secondly, because message types were categorized according to perceived hurt, the ceiling effect may reflect the fact that perceived hurt does not always correspond directly with perceived aggression. For example, message types such as "Swearing" and "Competence Attack" can be judged as equally aggressive despite differing in perceived hurt. Together, these factors suggest that while varying message content based on perceived hurt helps participants distinguish lower levels of aggression, it may be insufficient for differentiating the higher levels.

Perceptual inconsistency was also notable under the *Slightly Aggressive* condition for Language, likely due to the indirect nature of passive aggression, such as teasing or implied threats. These expressions are often open to interpretation [90], contributing to variability in how aggression is perceived. Messages were frequently misidentified as competence-related attacks, even within the *Not Aggressive* condition (5.26%). This may reflect heightened sensitivity to perceived criticism in customer service contexts, where statements that challenge competence can be interpreted as aggressive, even when unintended. These findings suggest that while Language can convey subtle forms of aggression, perceptions of such low-intensity, indirect aggression are highly dependent on individual interpretation.

Body movement and facial expressions proved especially effective in providing distinct and consistent gradations of perceived aggression across all levels. The application of Laban Movement Analysis factors, and Facial Action Unit Combinations resulted in clear differences in perceived aggression from *Not Aggressive* to *Very Aggressive*. These findings align with existing research on non-verbal communication, which emphasizes the universality and intuitiveness of both face and body cues for emotional and social signals [4, 55, 56]. Participants interpreted physical expressions in body cues, such as movement patterns and flow, more reliably than subtle message types in language or subtle variations in intensity and spectral balance in voice, demonstrating that non-verbal cues, provide readily interpretable and consistent markers of aggression, making them a highly reliable modality for aggression perception.

Overall, participants were sensitive to the systematically scaled aggression cues within each modality, indicating that our encoding of aggression intensity was meaningful and interpretable. This validates the robustness of our design while also revealing modality-specific strengths and weaknesses. The variability observed in linguistic aggression highlights the challenge of relying solely on verbal content to simulate social tension, particularly in cross-cultural or high-stakes service contexts. By contrast, the reliability of vocal and bodily cues suggests that multimodal IVAs should prioritize embodied aggression signals when aiming for perceptual clarity. Nevertheless, language remains critical for conveying subtle, indirect aggression, such as passive-aggressive remarks, which may be especially relevant in service interactions. Together, these insights motivate multimodal integration (tested in Experiment 2), where the strengths of each modality can compensate for the weaknesses of others, producing more realistic and interpretable simulations of aggression.

5 Experiment 2 (Multimodal Perception)

Prior work has explored multimodal emotional expression, but it remains unclear how individual modalities interact to shape aggression perception. Experiment 1 (Unimodal Perception) examined modality-specific effects, while Experiment 2 investigated how Language, Voice, and Body cues integrate and potentially interact when combined.

Originally, four aggression levels (*Not Aggressive*, *Slightly Aggressive*, *Aggressive*, *Very Aggressive*) were used. Based on Experiment 1, participants could not reliably distinguish *Aggressive* from *Very Aggressive* in Language cues, so we removed the *Aggressive* condition and focused on the three perceptually distinct levels, renamed *Low*, *Mid*, and *High*, to improve interpretability.

This experiment aimed to understand how participants perceive aggression when Language, Voice, and Body cues are combined. We asked: (1) How do participants perceive aggression when multiple cues are presented together—as a simple combination of independent cues, or does the combination feel stronger or weaker than expected? (2) Can participants reliably perceive aggression as meaningful and distinct when a virtual agent conveys it using integrated language, voice, body, and facial cues?

5.1 Participants

Experiment 2 included 19 participants (13 female, 6 male), aged 22 to 55 years (M = 30.6, SD = 9.0) with 2 to 28 years of professional experience (M = 7.3, SD = 7.0), reflecting the typical demographic and experience distribution of cabin crew available on commercial flights operated by our airline partner. This was a separate group of participants from Experiment 1. Recruitment procedures and compensation were consistent with those described in Experiment 1. All experimental protocols were approved by the Institutional Review Board (IRB).

5.2 Experimental Design

The multimodality experiment investigated how aggression is perceived when cues from Language, Voice, and Body (body movement & facial expression) were presented in combination. A within-subject design was used, with aggression level in each modality

as an independent variable. Each participant was exposed to 27 conditions (3 x 3 x 3), corresponding to all possible combinations of three aggression levels in Language *(L1–L3)*, three in Voice *(V1–V3)*, and three in Body *(B1–B3)*. For clarity, combinations are denoted using shorthand notation such as *L1V2B3*, which represents *Not Aggressive* Language *(L1)*, *Slightly Aggressive* Voice *(V2)*, and *Very Aggressive* Body *(B3)*.

The order of conditions was fully counterbalanced across participants to control for potential order effects. Participants' subjective ratings of perceived aggression served as the dependent variable.

5.3 Apparatus and Materials

To evaluate aggression perception across combined modalities, stimuli for Language, Voice, and Body were drawn from Experiment 1 and presented together through a Meta Quest Pro VR headset. This setup allowed participants to experience all modalities simultaneously in an immersive environment. Language phrases were delivered in spoken form rather than written text to enable natural integration with the voice samples.

The stimuli were synchronized using a networked Unity3D application (see Sec. 3.2), which ensured precise timing and consistent delivery. A researcher managed the presentation throughout the study to maintain experimental consistency and provide participants with a seamless, immersive experience.

5.4 Procedure

After providing informed consent, participants completed the multimodality experiment following a structured procedure:

(1) They completed the same briefing and familiarization steps as in Experiment 1.
(2) Participants were presented with stimuli representing all 27 combinations of aggression levels across Language, Voice, and Body, in a counterbalanced order (see Sec. 5.3 for details).
(3) After each stimulus, participants completed a questionnaire rating perceived aggression.

To manage the increased number of trials, the experiment was divided into three blocks of nine trials each, with a 5-minute break between blocks to reduce fatigue from prolonged VR headset use. The entire session lasted approximately 45 minutes per participant.

5.5 Measures & Hypotheses

5.5.1 Perceived Aggression. Participants rated each stimulus on a 5-point Likert scale (1 = Least Aggressive, 5 = Most Aggressive) immediately after exposure. This measure captured participants' subjective perception of aggression for each multimodal combination, allowing us to evaluate how cues from Language, Voice, and Body jointly influenced perceived aggression.

H1: Perceived Aggression in Multimodal Conditions. We hypothesize that when participants observe combined Language, Voice, and Body cues, they will perceive aggression as a balanced average of the emotional signals across modalities, rather than as a sum or any other non-linear combination that might amplify or dampen the overall impression.

H2: Emergence of Distinct Multimodal Aggression Levels. We hypothesize that participants' perceived aggression ratings across the 27 multimodal combinations (3 levels each of Language, Voice, and Body) will cluster into three perceptually distinct levels (*Low*, *Mid*, *High*). Each modality was systematically manipulated to produce three distinct levels of aggression, and we expect that their combination will preserve a meaningful and distinguishable aggression gradient.

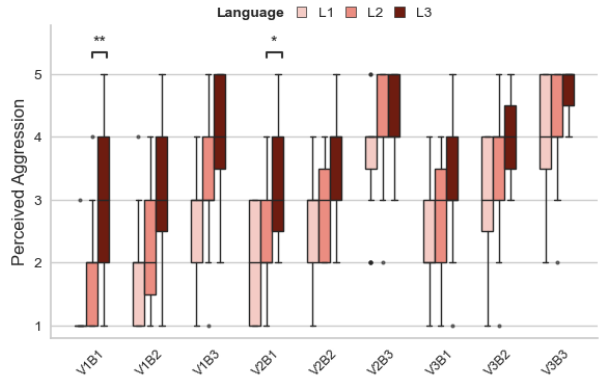
5.5.2 Qualitative Feedback. After the session, participants completed an open-ended survey about their perceptions of the aggressive IVA (full questions in Appendix A.2). They were prompted to elaborate on their reasoning and experiences, including which modality was most helpful in judging aggression, how they resolved conflicts between cues, and any uncertainty or difficulty involved in these judgments.

5.6 Results

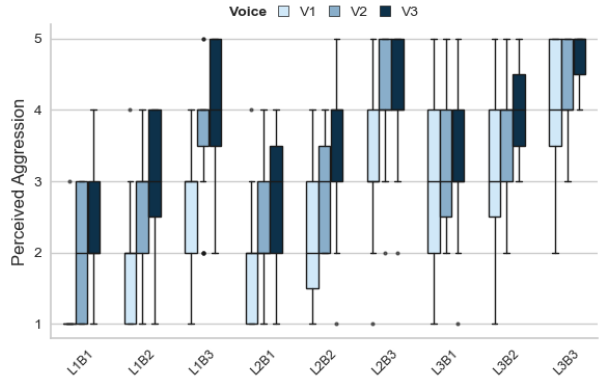
5.6.1 Perceived Aggression. The effects of the three modalities (Language, Voice, and Body) on perceived aggression in multimodal conditions were analysed. To evaluate the normality assumption for the three-way repeated measures ANOVA, the Shapiro-Wilk test was applied to the residuals (refer to Appendix A.3), which revealed a significant deviation from normality ($p < 0.01$). As such, the aligned rank transform (ART) method [93] was used, followed by a three-way repeated measures ANOVA on the transformed data. For any significant main effects or interactions, post hoc pairwise comparisons were conducted using ART contrasts with Tukey's HSD adjustment.

Main Effects of Language, Voice, and Body. Language had a significant main effect on perceived aggression ($F(2, 468) = 100.19$, $p < 0.001$, $\eta_p^2 = 0.30$). Participants perceived increasing aggression across Language levels, with mean ratings rising from *L1* ($M = 2.58$, $SD = 1.41$) to *L2* ($M = 2.99$, $SD = 1.38$) and *L3* ($M = 3.70$, $SD = 1.12$). Post hoc tests confirmed that each successive level was perceived as significantly more aggressive than the previous one (*L1-L2*: $p < 0.0001$; *L2-L3*: $p < 0.0001$; *L1-L3*: $p < 0.0001$). When examining the effects across specific Voice and Body conditions (without collapsing across these factors), *L3* was rated significantly higher than *L1*, particularly at lower Voice levels. Significant differences were observed in the *V1B2* and *V2B2* conditions ($p < 0.01$), as well as in the *V1B1*, *V1B3*, and *V2B1* conditions ($p < 0.001$). At higher Voice levels, these differences diminished, with only the *V3B2* condition showing a significant difference ($p < 0.05$). Significant differences were observed between *L2* and *L3* only for the V1B1 condition ($p < 0.01$) and the V2B1 condition ($p < 0.05$). No significant differences were observed between *L2* and *L1*.

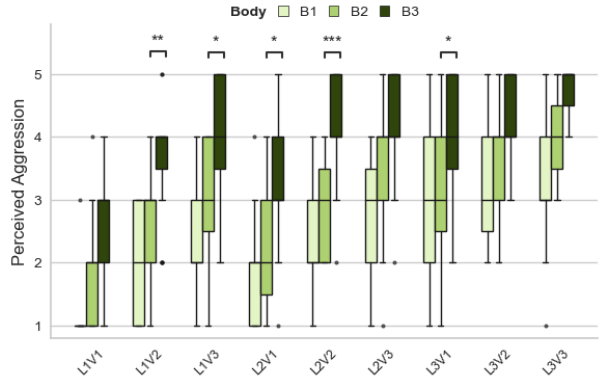
Voice also had a significant main effect on perceived aggression ($F(2, 468) = 65.98$, $p < 0.001$, $\eta_p^2 = 0.22$). Mean perceived aggression increased steadily with Voice intensity, from *V1* ($M = 2.56$, $SD = 1.51$) to *V2* ($M = 3.20$, $SD = 1.28$) and *V3* ($M = 3.51$, $SD = 1.26$). Post hoc analyses confirmed that participants perceived each increase as statistically significant (*V1-V2*: $p < 0.0001$; *V2-V3*: $p < 0.01$; *V1-V3*: $p < 0.0001$). This effect was most notable at lower Language levels when analysing across Language and Body conditions, with significant differences between *V1* and *V3* observed in conditions *L1B1* and *L2B1* ($p < 0.05$), *L1B1* and *L2B3* ($p < 0.01$), and *L1B3* and *L2B2* ($p < 0.001$). At the highest Language level, significant differences between *V1* and *V3* were found only in the *L3B2* condition ($p < 0.001$). Across all conditions, no significant



(a) Box plots showing perceived aggression ratings across three Language levels for each Voice-Body condition



(b) Box plots showing perceived aggression ratings across three voice levels for each Language-Body condition



(c) Box plots showing perceived aggression ratings across three Body levels for each Language-Voice condition

Figure 6: Box plots showing perceived aggression ratings across varying levels of Language, Voice, and Body modalities under different conditions.

differences were observed between *V1* and *V2*, or between *V2* and *V3*.

Lastly, Body showed a significant main effect on perceived aggression, having the largest effect size among the three modalities ($F(2, 468) = 158.05$, $p < 0.001$, $\eta_p^2 = 0.40$). Participants reliably distinguished between successive Body aggression levels, with mean ratings increasing from *B1* ($M = 2.42$, $SD = 1.21$), *B2* ($M = 2.97$, $SD = 1.16$), and *B3* ($M = 3.88$, $SD = 1.10$). All pairwise comparisons were significant (*B1-B2*: $p < 0.0001$; *B2-B3*: $p < 0.0001$; *B1-B3*: $p < 0.0001$). Unlike Language and Voice, Body aggression exhibited a more uniform effect across Language and Voice conditions. *B3* was rated significantly higher than *B1* across all conditions ($p < 0.001$), with weaker significance observed only in *L3V1* and *L3V2* ($p < 0.01$). Significant differences between *B3* and *B2* were observed in *L1V3*, *L2V1*, and *L3V1* ($p < 0.05$), as well as in *L1V2* ($p < 0.01$) and *L2V2* ($p < 0.001$). In contrast, no significant differences were found between *B2* and *B1* across all conditions.

Interaction Effects Among Language, Voice, and Body. A significant interaction was found between Language and Voice on perceived aggression ($F(4, 468) = 4.15$, $p = 0.002$, $\eta_p^2 = 0.034$), indicating that the effect of Language on perceived aggression varied depending on Voice levels. A difference-in-differences analysis indicated that disparity in perceived aggression between *L1* and *L3* was more pronounced at the lowest Voice level *V1* and these differences diminished as Voice levels increased to *V2* and *V3*. Specifically, a significant reduction in the *L1-L3* difference was observed when comparing *V1* with *V2* ($p < 0.001$) and *V1* with *V3* ($p < 0.0001$). Conversely, the effect of Voice on perceived aggression was also dependent on Language levels. Increase in perceived aggression from *V1* to *V3* was significantly reduced when comparing *L2* with *L3* ($p < 0.01$) and *L1* with *L3* ($p < 0.001$).

No significant interactions were found between Voice and Body ($F(4, 468) = 0.67$, $p = 0.611$, $\eta_p^2 = 0.006$), nor between Language and Body ($F(4, 468) = 2.14$, $p = 0.075$, $\eta_p^2 = 0.018$). Additionally, the three-way interaction among Language, Voice, and Body was non-significant ($F(8, 468) = 0.87$, $p = 0.538$, $\eta_p^2 = 0.015$).

The detailed pairwise comparisons showing the effects of Language, Voice, and Body on perceived aggression are provided in Table 6. Corresponding box plots of perceived aggression ratings across all three modalities are presented in Fig. 6, while Fig. 7 presents an interaction plot summarizing the combined effects of the three modalities on perceived aggression.

5.6.2 K-means Clustering Results. To evaluate whether the 27 combinations of Language, Voice, and Body levels produced perceptually distinct aggression levels, we first aggregated participants' ratings for each multimodal combination and then conducted a K-means clustering analysis implemented in the *scikit-learn* library (Python). Such an approach has been used in human perception research [15] to minimize within-cluster differences and maximize between-cluster differences in subjective ratings. We determined the optimal number of clusters using the elbow method, which indicated that three clusters best represented the data. Subsequently, we specified the number of clusters as 3 and set the random state to 42 to ensure reproducibility, while all other parameters were kept at their default values. The algorithm identified three clusters corresponding to *Low* (Cluster 0), *Mid* (Cluster 1), and *High* (Cluster 2) perceived aggression, as shown in Table 7 and visualized in Figure 8. A Kruskal-Wallis test confirmed a significant effect of the

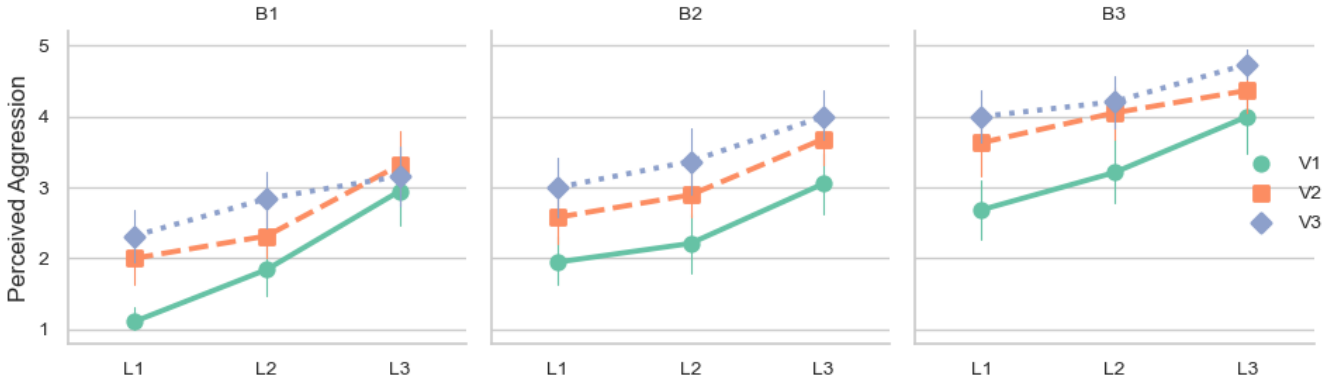


Figure 7: Interaction plot showing the combined effects of multimodal cues on perceived aggression. Language levels are L1, L2, and L3; Voice levels are V1, V2, and V3; Body levels are B1, B2, and B3. Points represent mean perceived aggression, and error bars show 95% confidence intervals, illustrating how the three cues interact to influence perception.

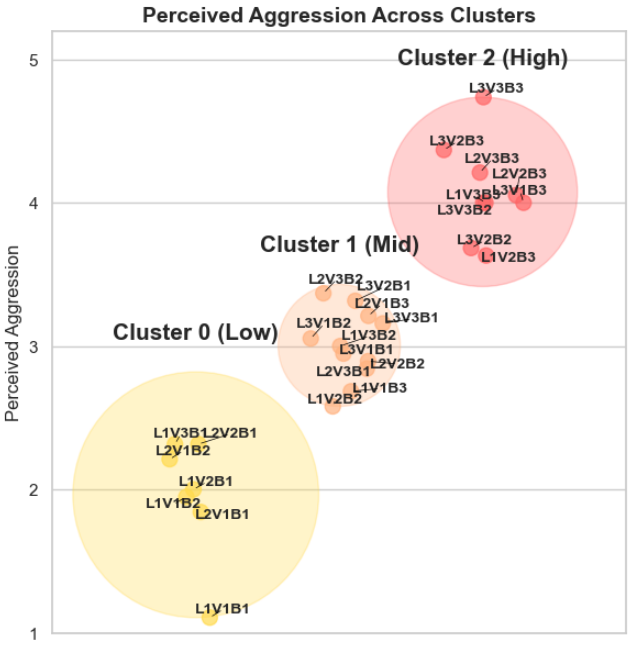


Figure 8: Scatter plot showing multimodal cue combinations, with each point representing the mean perceived aggression for that combination, color-coded by K-means clusters (Low, Mid, High). Shaded regions around cluster centroids indicate cluster spread, defined by the maximum intra-cluster distance.

cluster assignments on perceived aggression ratings (H = 223.707, $p < 0.0001$, $\eta^2 = 0.563$), and post hoc pairwise comparisons using Mann-Whitney U tests with Holm-Bonferroni correction showed significant differences between all clusters (Cluster 0–1: $p < 0.0001$, Cluster 1–2: $p < 0.0001$, Cluster 0–2: $p < 0.0001$).

To understand why these clusters emerged, we examined participants' ratings across all multimodal combinations. These clusters appear to result from perceptual averaging, where participants integrated cues from Language, Voice, and Body into a single percept rather than treating each cue in isolation. However, the strongest differences arose when aggression was scaled cohesively across multiple modalities, as the aligned signals reinforced one another and amplified the perceived distinctions between levels. Post hoc tests confirmed this pattern: while single-modality increases produced only occasional significant differences, dual-modality scaling yielded clearer and more consistent effects. For example, when Language and Body were scaled together, significant differences were observed across all Voice conditions: under *V1*, *L1B1* vs *L2B2* ($p < 0.05$) and *L2B2* vs *L3B3* ($p < 0.0001$); under *V2*, *L2B2* vs *L3B3* ($p < 0.0001$); under *V3*, *L1B1* vs *L2B2* ($p < 0.01$) and *L2B2* vs *L3B3* ($p < 0.0001$). These effects were consistent across other dual-modality scaling conditions, especially when Body was involved, showing that including Body cues consistently strengthened perceived differences between aggression levels.

The largest effects were observed when all three modalities were scaled simultaneously. Across trio-modality levels, all pairwise comparisons yielded highly significant differences (e.g., *L1V1B1* vs. *L2V2B2*, $p < 0.0001$; *L2V2B2* vs. *L3V3B3*, $p < 0.0001$), representing the strongest effects in the study. Notably, the combinations *L1V1B1*, *L2V2B2*, and *L3V3B3*, corresponding to the *Low*, *Mid*, and *High* clusters, also showed the lowest perceptual variability within each cluster (SD = 0.46, 0.81, 0.45, respectively), indicating that participants consistently perceived these levels as distinct. These results suggest that while perceptual averaging across modalities contributed to cluster formation, the clearest and most reliable distinctions emerged when aggression cues in Language, Voice, and Body were scaled together, reinforcing one another to produce a strong, unified perceptual signal.

Fig. 9 presents box plots showing these ratings, highlighting how dual- and trio-modality scaling affects perceived aggression relative to single-modality scaling in Fig. 6. These plots highlight

Table 6: Pairwise comparisons of perceived aggression ratings, represented by mean (SD), across the three modalities—Language, Voice, and Body—evaluated at different levels of aggression under varying conditions. A '∼ symbol indicates no significant difference between levels, while '<' denotes a significant difference. Significance levels are denoted as follows: $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

(a) Pairwise comparisons across Language aggression levels, evaluated under varying Voice and Body conditions.

Condition (V-B)	Mean (SD)			Comparison Results
	L1	*L2*	*L3*	
V1B1	1.11 (0.46)	1.84 (0.90)	2.95 (1.08)	*L1 ∼ L2, L2 < L3**, L1 < L3****
V1B2	1.95 (0.85)	2.21 (0.92)	3.05 (1.00)	*L1 ∼ L2, L2 ∼ L3, L1 < L3***
V1B3	2.68 (1.00)	3.21 (1.03)	4.00 (1.11)	*L1 ∼ L2, L2 ∼ L3, L1 < L3****
V2B1	2.00 (0.82)	2.32 (0.82)	3.32 (1.06)	*L1 ∼ L2, L2 < L3*, L1 < L3****
V2B2	2.58 (0.84)	2.89 (0.81)	3.68 (0.82)	*L1 ∼ L2, L2 ∼ L3, L1 < L3***
V2B3	3.63 (1.00)	4.05 (0.85)	4.37 (0.68)	*L1 ∼ L2, L2 ∼ L3, L1 ∼ L3*
V3B1	2.32 (0.82)	2.84 (0.96)	3.16 (0.96)	*L1 ∼ L2, L2 ∼ L3, L1 ∼ L3*
V3B2	3.00 (1.00)	3.37 (1.07)	4.00 (0.75)	*L1 ∼ L2, L2 ∼ L3, L1 < L3**
V3B3	4.00 (0.88)	4.21 (0.85)	4.74 (0.45)	*L1 ∼ L2, L2 ∼ L3, L1 ∼ L3*

(b) Pairwise comparisons across Voice aggression levels, evaluated under varying Language and Body conditions.

Condition (L-B)	Mean (SD)			Comparison Results
	V1	*V2*	*V3*	
L1B1	1.11 (0.46)	2.00 (0.82)	2.32 (0.82)	*V1 ∼ V2, V2 ∼ V3, V1 < V3**
L1B2	1.95 (0.85)	2.58 (0.84)	3.00 (1.00)	*V1 ∼ V2, V2 ∼ V3, V1 < V3***
L1B3	2.68 (1.00)	3.63 (0.96)	4.00 (0.88)	*V1 ∼ V2, V2 ∼ V3, V1 < V3****
L2B1	1.84 (0.90)	2.32 (0.82)	2.84 (0.96)	*V1 ∼ V2, V2 ∼ V3, V1 < V3**
L2B2	2.21 (0.92)	2.89 (0.81)	3.37 (1.07)	*V1 ∼ V2, V2 ∼ V3, V1 <V3****
L2B3	3.21 (1.03)	4.05 (0.85)	4.21 (0.85)	*V1 ∼ V2, V2 ∼ V3, V1 < V3***
L3B1	2.95 (1.08)	3.32 (1.06)	3.16 (0.96)	*V1 ∼ V2, V2 ∼ V3, V1 ∼ V3*
L3B2	3.05 (0.97)	3.68 (0.82)	4.00 (0.75)	*V1 ∼ V2, V2 ∼ V3, V1 < V3**
L3B3	4.00 (1.11)	4.37 (0.68)	4.74 (0.45)	*V1 ∼ V2, V2 ∼ V3, V1 ∼ V3*

(c) Pairwise comparisons across Body aggression levels, evaluated under varying Language and Voice conditions.

Condition (L-V)	Mean (SD)			Comparison Results
	B1	*B2*	*B3*	
L1V1	1.11 (0.46)	1.95 (0.85)	2.68 (1.00)	*B1 ∼ B2, B2 ∼ B3, B1 < B3****
L1V2	2.00 (0.82)	2.58 (0.84)	3.63 (0.96)	*B1 ∼ B2, B2 < B3**, B1 < B3****
L1V3	2.32 (0.82)	3.00 (1.00)	4.00 (0.88)	*B1 ∼ B2, B2 < B3*, B1 < B3****
L2V1	1.84 (0.90)	2.21 (0.92)	3.21 (1.03)	*B1 ∼ B2, B2 < B3*, B1 < B3****
L2V2	2.32 (0.82)	2.89 (0.81)	4.05 (0.85)	*B1 ∼ B2, B2 < B3***, B1 < B3****
L2V3	2.84 (0.96)	3.37 (1.07)	4.21 (0.85)	*B1 ∼ B2, B2 ∼ B3, B1 < B3****
L3V1	2.95 (1.08)	3.05 (0.97)	4.00 (1.11)	*B1 ∼ B2, B2 < B3*, B1 < B3***
L3V2	3.32 (1.06)	3.68 (0.82)	4.37 (0.68)	*B1 ∼ B2, B2 ∼ B3, B1 < B3***
L3V3	3.16 (0.96)	4.00 (0.75)	4.74 (0.45)	*B1 ∼ B2, B2 ∼ B3, B1 < B3****

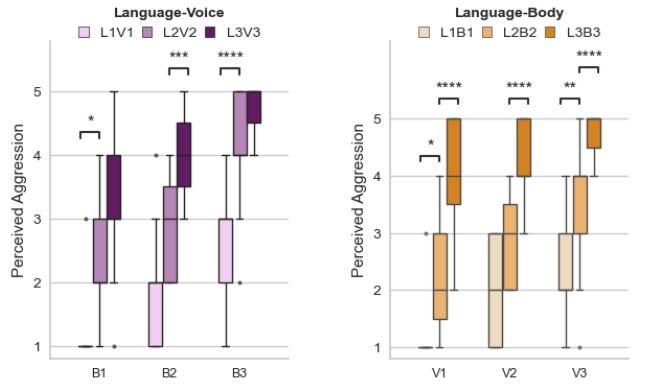
Table 7: K-means clustering results showing multimodal combinations grouped into clusters with their corresponding perceived aggression ratings.

Cluster ID	Multimodal Combinations	Perceived Aggression Mean (SD)
0	*L1V1B1, L1V1B2, L1V2B1, L1V3B1, L2V1B1, L2V1B2, L2V2B1*	1.96 (0.42)
1	*L1V1B3, L1V2B2, L1V3B2, L2V1B3, L2V2B2, L2V3B1, L2V3B2, L3V1B1, L3V1B2, L3V2B1, L3V3B1*	3.00 (0.25)
2	*L1V2B3, L1V3B3, L2V2B3, L2V3B3, L3V1B3, L3V2B2, L3V2B3, L3V3B2, L3V3B3*	4.08 (0.33)

how combining and scaling modalities cohesively produces clear differences in perceived aggression.

5.7 Discussion

H1: Perceived Aggression in Multimodal Conditions. We hypothesized that when multiple modalities are combined, participants would integrate cues independently, effectively averaging aggression signals across Language, Voice, and Body rather than summing or amplifying them. Our findings generally support this hypothesis. Participants' ratings typically reflected an averaging process: when one modality was highly aggressive and another less so, the perceived aggression tended to fall between the two extremes rather than being dominated by the most intense cue. This suggests that integration is generally balanced and independent across modalities. However, we observed a specific interaction between Language and Voice. In some cases, high aggression in one modality appeared to dampen the perceived impact of the other, indicating subtle non-linear effects. Hence, while averaging predominates in multimodal perception, interactions between closely



(a) Perceived aggression when Language and Voice are scaled together

(b) Perceived aggression when Language and Body are scaled together

(c) Perceived aggression when Voice and Body are scaled together

(d) Perceived aggression when Language, Voice and Body are scaled together

Figure 9: Box plots showing perceived aggression ratings when two modalities are scaled together: (a) Language and Voice, (b) Language and Body, (c) Voice and Body, and when all three modalities are scaled together: (d) Language, Voice, and Body.

related modalities, such as Language and Voice, can slightly modify this pattern.

Body consistently emerged as the strongest predictor of perceived aggression, supported by both quantitative and qualitative evidence. Statistically, Body cues produced the clearest separation and most linear progression across aggression levels, with the largest effect size. In the post-session survey, 15 of 19 participants reported relying primarily on Body to judge aggression. For example, P2 noted, "I chose which showed me more aggression," suggesting that Body cues were particularly salient and noticeable, perhaps because they are visual and therefore more immediately perceptible. These findings underscore the importance of non-verbal signals in shaping participants' perceptions of aggression.

Despite the prominence of body cues, participants did not perceive them in isolation. When Body signalled high aggression but Language or Voice indicated neutrality, ratings tended to reflect intermediate aggression levels. This suggests that participants integrate information across modalities, balancing conflicting cues rather than defaulting to the most intense signal. One participant, P10, remarked, "If the person's behaviour is typically non-aggressive but is using harsh words, you might interpret it as situational frustration rather than true aggression." Similar reflections were reported by other participants, indicating a common tendency to weigh and average information from multiple channels when cues conflict.

We also found the extent to which participants averaged emotional information across modalities depended on whether the cues were congruent or incongruent. For fully congruent combinations (i.e., Language, Voice, and Body all at the same aggression level), the averaging effect was less pronounced because all cues conveyed a similar level of aggression. These combinations (*L1V1B1*, *L2V2B2*, *L3V3B3*) also exhibited the lowest variability, indicating that aligned cues reinforce one another to form a stable, unified percept rather than dilute participants' perception of aggression.

In contrast, fully or partially incongruent combinations (i.e., Language, Voice, and Body at different aggression levels, such as *L1V3B2*) elicited a stronger averaging effect. When cues conflicted, participants blended information across modalities, producing intermediate and less distinct aggression ratings. Integration of cues was also not uniform; participants drew on the modalities in different ways, resulting in more subjective and variable interpretations of the same incongruent stimuli. The conflicting cues introduced substantial ambiguity, as reflected in participants' accounts. For example, P8 remarked, "When a passenger uses strong tones but polite language, it makes it difficult for me to determine whether they are genuinely angry or just have a habitual way of speaking." Similarly, P10 noted, "When one modality, like language, conveys aggression but another, like tone of voice, is calm, it's hard to determine the speaker's true intention." These comments illustrate how incongruent cues increase uncertainty and hinder participants' ability to infer the intended aggression level.

Although our study did not directly measure cognitive workload (e.g., via NASA-TLX) to assess user effort, the pattern of results suggests that multimodal congruence supports more reliable perception of aggression, echoing foundational studies demonstrating that alignment across channels is essential for interpretable and reliable emotional communication in virtual agents [18, 23, 44]. Participants showed the greatest agreement when all modalities were aligned (*L1V1B1*, *L2V2B2*, *L3V3B3*), while conflicting cues produced perceptual instability. Prior research has also shown that cross-modal incongruence reduces observers' accuracy and impairs emotion recognition, highlighting the additional processing demands placed on the perceiver when attempting to reconcile conflicting cues [88]. This suggests that aligning cues across modalities is important for designing aggressive IVAs, as congruent signals create clearer, more interpretable levels of aggression that trainees can effectively perceive and respond to.

Despite this, incongruent cues remain valuable for training, as verbal, vocal, and bodily signals are frequently mismatched in everyday interactions (e.g., sarcasm, polite words delivered with an irritable tone, or conflicting verbal and body language) [37]. Exposing trainees to such cues helps them practice interpreting mixed

emotional signals, which is especially important in cross-cultural contexts where norms for expressing emotion differ.

Overall, our results align with prior work on multimodal emotion perception [92, 94], which shows that emotional signals are perceptually blended rather than strictly hierarchical. Our results highlight the central role of body movement and facial expressions in perceiving aggression, while showing that no single modality operates in isolation. For effective aggression modelling in virtual agents, non-verbal cues should remain central, but always considered within the broader context of multimodal integration.

H2: Emergence of Distinct Multimodal Aggression Levels. We hypothesized that participants' perceived aggression across the 27 multimodal combinations (three levels each of Language, Voice, and Body) would form three distinct levels corresponding to *Low*, *Mid*, and *High* aggression. Our clustering results support this hypothesis, revealing three well-separated levels that reflect meaningful gradations in perceived aggression. However, the clarity of this gradient depended strongly on which set of combinations were used: combinations that differed only subtly between levels produced weaker distinctions, indicating that adjacent levels must be sufficiently different for participants to perceive them as separate.

Manipulating a single modality in isolation rarely produced noticeable changes in perceived aggression. Body cues had slightly more influence than Language or Voice when varied alone, but none independently drove consistent shifts in perception. More pronounced differences emerged when two modalities increased together, particularly Body and Language, suggesting that physical and linguistic cues can jointly shape aggression perception. The clearest and most reliable level distinctions occurred only when all three modalities were scaled simultaneously, creating a strong, monotonic progression from low to high aggression.

This pattern indicates that fully aligned, congruent cue combinations are the most effective for defining discrete aggression levels: *L1V1B1* for *Low*, *L2V2B2* for *Mid*, and *L3V3B3* for *High*. These aligned combinations span the full dynamic range that participants can perceive from the IVA, providing clear anchors at the lower, midpoint, and upper ends of the aggression scale. Their high contrast ensures that each level is reliably distinguishable and that the overall progression remains interpretable and perceptually salient.

As highlighted in H1, incongruent cue combinations, where modalities signal different intensities, naturally produce intermediate aggression ratings due to perceptual averaging. Rather than forming clean categories, these combinations occupy the spaces between the three aligned anchors. This makes them useful for generating subtle, fine-grained variations in aggressive behaviour, though doing so requires careful calibration to ensure that mixed signals remain interpretable across adjacent levels.

From a training perspective, the choice between using congruent or incongruent cues depends on the learning objective. Fully aligned combinations provide high-contrast, easily recognisable shifts and are therefore well-suited for establishing baseline behaviours or helping trainees reliably distinguish among Low, Mid, and High aggression. In contrast, incongruent or partially congruent combinations support training in perceptual sensitivity, as they introduce more nuanced, harder-to-detect shifts that mirror the ambiguity often encountered in real-world interpersonal interactions. By integrating both types of cue combinations, the IVA can support training scenarios that require either clear categorical transitions or more subtle, perceptually challenging variations in aggressive behaviour.

More importantly, because our IVA can express three perceptually distinct aggression levels (*Low*, *Mid*, *High*) rather than a simple binary on–off state, it enables a more granular modulation of aggressive behaviour in real time based on user feedback or contextual input. This supports emotion-adaptive training systems, allowing the IVA to adjust its internal aggression state dynamically in response to trainee actions, stress levels, or specific learning objectives [48, 60, 87]. By adapting its multimodal cues as interactions unfold, the IVA can produce more realistic and context-sensitive behaviour that reflects the variability and complexity of real-world human aggression, thereby enhancing both training fidelity and pedagogical value [96].

Although IVAs can approximate key aspects of human social interaction, empirical work continues to document limitations in behavioural realism (such as unnatural motion, reduced non-verbal bandwidth, and synthetic voice) that can diminish social presence and emotional impact relative to real-human encounters [26, 53]. These constraints make it difficult for digital agents to reproduce the full intensity and complexity characteristic of aggressive behaviour [9]. Incorporating cues from multiple modalities may attenuate some of these shortcomings by enhancing perceived realism and immersion [66]. Nevertheless, the fidelity gap constrains how closely IVAs can emulate real-world aggressive encounters and may influence the extent to which skills learned in IVA-based environments transfer to contexts involving higher-intensity or more severe aggression. IVAs may thus be best positioned as a complementary modality that augments, rather than replaces, in-person scenario training. Section 7 provides further discussion of generalizability limitations, realism constraints, and potential novelty effects.

In summary, our results show that the three aggression levels defined for each modality (*Low*, *Mid*, *High*) are preserved when combined, indicating that the multimodal aggression spectrum remains interpretable. However, the clarity of this gradient depends on the specific cue combinations used to represent each level. Participants perceived changes in aggression most accurately when cues were coordinated and incrementally scaled across all modalities. Because aggression perception is inherently multimodal, aligning cues across Language, Voice, and Body is essential to produce meaningful shifts in perceived aggression. While body and facial expressions contribute greatly to aggression perception, they are not sufficient on their own. These findings underscore the importance of systematically testing multimodal cue combinations when designing aggressive behaviour for IVAs, as different configurations can substantially affect how clearly users perceive changes in aggression.

6 Design Guidelines

Based on comprehensive findings from our experiments, we propose a set of design guidelines to support the creation of intelligent virtual agents (IVAs) that effectively communicate aggressive intent. These guidelines provide actionable insights for researchers and designers aiming to create IVAs that engage users authentically,

improve emotional recognition, and enhance interaction outcomes. While grounded in the customer service domain, these principles can also guide the design of emotionally expressive virtual agents in other domains, such as healthcare training, law enforcement simulations, or gaming, that aim to convey emotions along a continuum.

(1) **Design IVAs with Multi-Modal, Multi-Level Aggression.** To effectively convey graded aggression in a multimodal context, designers should systematically scale the different modalities involved, and iteratively test user perception to ensure that the aggression levels are clearly distinguishable. Participants' ratings from our multimodal study reliably clustered into three distinct aggression levels (Low, Mid, High) after systematic testing across different multimodality combinations (see Table 7), demonstrating that our IVA can convey multimodal aggression along a spectrum and support real-time modulation of its behaviour.

(2) **Apply Validated Aggression Parameters for each Modality.** Guided by prior literature on aggression expression, our IVA incorporates parameters expected to convey graded aggression, which our experiments confirm are perceptible to participants across unimodal and multimodal contexts. Designers can therefore adopt the validated aggression parameters from the 2 studies as a baseline for each modality:
   - **Language:** Message type categorisation (see Table 1)
   - **Voice:** Style degree parameters (see Table 2)
   - **Body Movement and Facial Expressions:** LMA and Effort-Shape qualities (see Table 3); Facial AU combination intensities (see Fig. 3)

   Designers are encouraged to adjust these parameters further to improve gradations, as not all participants in our unimodality experiment reliably perceived differences in the manipulated cues, particularly for more subtle attributes.

(3) **Prioritize Non-Verbal Cues for Clear Aggression Signalling.** Our multimodal perception study showed that body movements and facial expressions were the most effective modality in conveying distinct and consistent gradations of perceived aggression, with the largest effect sizes across levels. Participants reported relying on body cues more than language or voice to interpret aggression, and they interpreted physical expressions, such as movement patterns and flow, more reliably than nuanced message types in language or subtle variations in intensity and spectral balance in voice. Based on these perception insights, designers should leverage structured frameworks such as Laban Movement Analysis (LMA) and Facial Action Unit (FAU) combinations to craft non-verbal cues that users can consistently perceive, ensuring clarity, interpretability, and reliability in aggression communication.

(4) **Coordinate Modalities to Support User Perception.** Based on our multimodal perception study, aggression is perceived most effectively when language, voice, body movement, and facial expression cues are aligned in intensity. Conflicting or disjointed cues can confuse users and lead to subjective interpretations, reducing their ability to interpret aggression reliably. While body movement and facial expressions often carry the strongest weight, designers should integrate all modalities holistically rather than relying on any single channel. Presenting a series of fully aligned cues, where aggression is scaled simultaneously across all modalities, helps users reliably perceive changes in graded aggression, enhancing emotional understanding and recognition. However, incongruent cues may still be relevant when applied deliberately and with clear instructional intent. Carefully selected incongruent combinations can expose trainees to the ambiguous signals commonly encountered in real-world interactions. Such use should be carefully controlled, as excessive or unstructured incongruence may undermine clarity and reduce training effectiveness.

(5) **Account for Ambiguity in Subtle Language.** Results from our unimodality perception study showed that passive-aggressive language (e.g., teasing or threats) was often ambiguous and open to interpretation, leading to higher variability in user perceptions at lower intensities of aggression. Perception depended heavily on context, tone, and the relationship between speaker and listener. Designers should pilot test subtle language cues to ensure they are understood as intended by target users.

7 Limitations And Future Work

Our approach has several limitations that warrant consideration and motivate important directions for future work.

Relationship between Perceived Hurt and Aggression. In the language modality, aggression levels were formulated using varying degrees of perceived hurt in message types. As a result, the relationship between perceived hurt and perceived aggression may be overstated, as other factors such as resilience and empathy can dampen this connection [95]. This variability may influence how different users perceive the aggressive language generated by the model.

Aggression Cues Based on Western Norms. Another limitation is that our findings may not generalize reliably to other cultural contexts. The model was developed using specific message types, AU intensities, vocal roughness, and movement parameters, which primarily reflect Western norms of anger expression. Cross-cultural differences in emotional display rules may therefore alter how users perceive aggression in IVAs, limiting the universality of the results.

Bias towards Reactive Aggression. The present model is biased toward particular psychological processes underlying aggression, relying on perceptual cues (language, voice, body movement, and facial expressions) as indicators. While these cues capture manifestations of emotional or impulsive aggression, they may insufficiently represent instrumental or goal-directed aggression, which can be enacted with minimal overt emotional expression [46]. Examples include intimidation or bullying for financial or political gain. Consequently, the model's applicability may be limited in contexts where intent-driven aggression is more relevant.

Continuity of Emotional Expression. Aggression in our model is expressed only during active speech and movement; idle periods are not animated. This may reduce the perceived continuity of emotional expression, potentially weakening the realism of the IVA and the consistency of perceived aggression.

Generalizability to High-Stakes Environments. This study focused on non-violent, reactive aggression typical in everyday customer-service interactions, such as aviation. It remains unclear whether the perceptual patterns observed would hold in higher-stakes, safety-critical environments, such as emergency care or public safety. Nevertheless, our work provides an initial empirical basis for understanding how users perceive multimodal aggression cues in virtual agents. Future research should explore whether these mechanisms generalize to other high-contact service contexts, such as retail, hospitality, or healthcare, and examine how variations in task demands, emotional intensity, and risk profiles influence aggression perception.

Understanding User Affective and Behavioural Responses. While this study focuses on users' ability to perceive and differentiate multimodal aggression cues, effective training systems also require an understanding of how such stimuli influence users' affective and behavioural responses, including stress, engagement, and de-escalation behaviour. One challenge is the potential for novelty effects, whereby trainees' heightened emotional responses during initial exposure to an aggressive IVA may attenuate over repeated interactions as familiarity increases [49]. Future work should therefore assess behavioural and physiological responses to aggressive IVAs across multiple training sessions, using measures such as heart rate, skin conductance, or pupil dilation. Such longitudinal evaluations would provide a more objective, real-time account of how stress, arousal, and engagement evolve over time, strengthening the practical relevance of multimodal aggression modelling for training contexts.

Expanding IVA Capability and Adaptive Modelling Further work can also be done to implement personalisation features, allowing the IVA to adjust its aggression output based on the user's sensitivity or past interactions, creating more tailored and effective experiences. Additionally, future research could extend the model to include physical aggression cues beyond body movement and facial expressions, such as dynamic postures, gestures, or approach behaviours, and incorporate haptic feedback to simulate push, pull, or shaking motions [34]. As violence represents an extreme subset of aggression, carefully integrating violent behaviours may also be important for capturing the full spectrum of aggressive expression.

8 Conclusion

In this paper, we systematically investigated user responses to an aggression simulation model across three modalities (language, voice, body movement and facial expressions) through two user-perception experiments. Experiment 1 examined each modality in isolation, while Experiment 2 explored their combined effects. Based on our findings, we propose design guidelines for creating intelligent virtual agents (IVAs) that convey graded aggression. Participants reliably perceived aggression across modalities, except for higher levels in language, which were harder to distinguish. When combined, modalities produced clear, three-level aggression distinctions (low, mid, high). These results support a multimodal aggression simulation framework for IVAs, which can enhance emotional resilience and problem-solving skills in training across desktop, mobile, virtual reality, and mixed-reality environments. We hope that it provides a foundation for future research on multimodal aggression in IVAs and inspires others to build virtual agents that can convey graded emotional states in training and interactive environments.

9 Acknowledgements on the Use of Generative AI

We used a large language model, specifically OpenAI's GPT-4, to generate neutral utterances and rephrase them to convey message types corresponding to intended aggression levels. Full details of this process are provided in Sec. 3.1.2. The authors take full responsibility for all outputs produced by the AI and for their use in this research.

Acknowledgments

This research is supported by National Research Foundation, Singapore and A*STAR, under its RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) grant call (Grant No. I2001E0059) – SIA-NUS Digital Aviation Corp Lab. The authors wish to express their sincere gratitude to Mr. James Chang, Ms. Pauline Koh and Ms. Samantha Yoon of Singapore Airlines, who oversee Cabin Crew Learning and Development, for their invaluable assistance in facilitating the user studies that informed this research. The authors also extend their thanks to Mr. Ng Yan Fei of the NUS Immersive Reality Lab for his contributions in developing the applications used in the study. They also extend their heartfelt appreciation to Ms. Liu Chang, Mr. Matthew Chan, and Mr. Cheng Hao Jie of the same lab for their essential assistance in conducting the user studies.

References

[1] Pengcheng An, Ziqi Zhou, Qing Liu, Yifei Yin, Linghao Du, Da-Yuan Huang, and Jian Zhao. 2022. Vibemoji: Exploring user-authoring multi-modal emoticons in Social Communication. *CHI Conference on Human Factors in Computing Systems* (Apr 2022), 1–17.

[2] Craig A. Anderson and Brad J. Bushman. 2002. Human aggression. *Annual Review of Psychology* 53, 1 (Feb 2002), 27–51.

[3] Horst Arndt and Richard W Janney. 1991. Verbal, prosodic, and kinesic emotive contrasts in speech. *Journal of pragmatics* 15, 6 (1991), 521–549.

[4] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 6111 (2012), 1225–1229.

[5] Rainer Banse and Klaus R Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology* 70, 3 (1996), 614.

[6] Tanja Bänziger, Georg Hosoya, and Klaus R Scherer. 2015. Path models of vocal emotion communication. *PloS one* 10, 9 (2015), e0136675.

[7] Christian Becker, Helmut Prendinger, Mitsuru Ishizuka, and Ipke Wachsmuth. 2005. Evaluating affective feedback of the 3D agent max in a competitive cards game. In *Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1*. Springer, 466–473.

[8] Leonard Berkowitz. 1989. Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin* 106, 1 (1989), 59–73.

[9] Romy Blankendaal, Tibor Bosse, Charlotte Gerritsen, Tessa de Jong, and Jeroen de Man. 2015. Are aggressive agents as scary as aggressive humans?. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 553–561.

[10] Gwen Bonner and Sue McLaughlin. 2007. The psychological impact of aggression on nursing staff. *British Journal of Nursing* 16, 13 (2007), 810–814.

[11] Tibor Bosse, Charlotte Gerritsen, and Jeroen de Man. 2015. Evaluation of a virtual training environment for aggression de-escalation. In *Proceedings of Game-On*. 48–58.

[12] Tibor Bosse, Tilo Hartmann, Romy AM Blankendaal, Nienke Dokter, Marco Otte, and LF Goedschalk. 2018. Virtually bad: A study on virtual agents that physically threaten human beings. In *AAMAS 2018 Conference: July 10-15, 2018 Stockholm*. ACM Press, 1258–1266.

[13] Roger Bougie, Rik Pieters, and Marcel Zeelenberg. 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the academy of marketing science* 31, 4 (2003), 377–393.

[14] D.C. Brogan, R.A. Metoyer, and J.K. Hodgins. 1998. Dynamically simulated characters in Virtual Environments. *IEEE Computer Graphics and Applications* 18 (1998), 58–69.

[15] Shaoyu Cai, Pingchuan Ke, Takuji Narumi, and Kening Zhu. 2020. Thermairglove: A pneumatic glove for thermal perception and material identification in virtual reality. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 248–257.

[16] Mario Callegaro, Michael H. Murakami, Ziv Tepman, and Vani Henderson. 2015. Yes–no answers versus check-all in self-administered modes: A systematic review and Analyses. *International Journal of Market Research* 57, 2 (Mar 2015), 203–224.

[17] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[18] Che-Jui Chang, Samuel S Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. The importance of multimodal emotion conditioning and affect consistency for embodied conversational agents. *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Mar 2023), 790–801.

[19] Pan Chen, Emil F. Coccaro, and Kristen C. Jacobson. 2012. Hostile attributional bias, negative emotional responding, and aggression in adults: Moderating effects of gender and impulsivity. *Aggressive Behavior* 38, 1 (Jan 2012), 47–63.

[20] Céline Clavel, Justine Plessier, Jean-Claude Martin, Laurent Ach, and Benoit Morel. 2009. Combining facial and postural expressions of emotions in a virtual character. In *Intelligent Virtual Agents: 9th International Conference, IVA 2009 Amsterdam, The Netherlands, September 14-16, 2009 Proceedings 9*. Springer, 287–300.

[21] Rodney M Coe. 1964. Conflict, interference, and aggression: computer simulation of a social process. *Behavioral science* 9, 2 (1964), 186.

[22] Elizabeth A Crane and M Melissa Gross. 2013. Effort-shape characteristics of emotion-related body movement. *Journal of Nonverbal Behavior* 37 (2013), 91–105.

[23] Chris Creed and Russell Beale. 2008. Psychological responses to simulated displays of mismatched emotional expressions. *Interacting with Computers* 20, 2 (2008), 225–239.

[24] Melanie Davis, Graham L. Bradley, Jason I. Racz, Samantha Ferguson, and Nicholas J. Buys. 2023. Understanding passenger hostility in transit: A systematic review. *Current Psychology* 43, 1 (Jan 2023), 132–154.

[25] John Dollard, Neal E. Miller, Leonard W. Doob, O. H. Mowrer, and Robert R. Sears. 1939. Frustration and aggression. *Frustration and aggression* (1939).

[26] Haoyang Du, Kiran Chhatre, Christopher Peters, Brian Keegan, Rachel McDonnell, and Cathy Ennis. 2025. Synthetically expressive: Evaluating gesture and voice for emotion and empathy in VR and 2D scenarios. *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents* (Sep 2025), 1–10.

[27] Paul Ekman. 1982. Methods for measuring facial action. *Handbook of methods in nonverbal behavior research* (1982), 45–90.

[28] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[29] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. 2002. *Facial Action Coding System: Manual and Investigator's Guide*. Research Nexus, Salt Lake City, UT.

[30] Elodie Etienne, Marion Ristorcelli, Sarah Saufnay, Aurélien Quilez, Rémy Casanova, Michael Schyns, and Magalie Ochs. 2024. A systematic review on the socio-affective perception of Ivas' multi-modal behaviour. *Proceedings of the ACM International Conference on Intelligent Virtual Agents* (Sep 2024), 1–10.

[31] Delphine Franco, Ruben Vanderlinde, and Martin Valcke. 2018. Dealing with aggression in the classroom: online clinical simulations to measure mastery of aggression management competences. In *INTED2018 Proceedings*. IATED, 9140–9148.

[32] Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. 2018. Spoken conversational ai in video games: Emotional dialogue management increases user engagement. In *Proceedings of the 18th international conference on intelligent virtual agents*. 179–184.

[33] G. GILBERT. 1972. Distance between sets. *Nature* 239, 5368 (Sep 1972), 174–174.

[34] Linford Goedschalk, Tibor Bosse, and Marco Otte. 2018. Get your virtual hands off me! – developing threatening ivas using haptic feedback. *Communications in Computer and Information Science* (2018), 61–75.

[35] Ed Groff. 1995. Laban movement analysis: Charting the ineffable domain of human movement. *Journal of Physical Education, Recreation & Dance* 66, 2 (1995), 27–30.

[36] Masum Hasan, Cengiz Ozel, Sammy Potter, and Ehsan Hoque. 2023. Sapien: Affective Virtual Agents powered by large language models*. *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (Sep 2023), 1–3.

[37] Cornelia Herbert. 2021. How to deal with incongruence? the role of social perception and bodily facial feedback in emotion recognition in human agent interaction – evidence from psychology as potential and challenge for multimodal user-centered approaches. *Communications in Computer and Information Science* (2021), 28–39.

[38] Viviane Herdel, Anastasia Kuzminykh, Andrea Hildebrandt, and Jessica R. Cauchard. 2021. Drone in love: Emotional perception of facial expressions on flying robots. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (May 2021), 1–20.

[39] Natalie Hube, Kresimir Vidackovic, and Michael Sedlmair. 2022. Using expressive avatars to increase emotion recognition: A pilot study. *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (Apr 2022), 1–7.

[40] Dominic A Infante. 1995. Teaching students to understand and control verbal aggression. *Communication Education* 44, 1 (1995), 51–63.

[41] Dominic A Infante, Teresa A Chandler, and Jill E Rudd. 1989. Test of an argumentative skill deficiency model of interspousal violence. *Communications monographs* 56, 2 (1989), 163–177.

[42] Dominic A Infante and Andrew S Rancer. 1996. Argumentativeness and verbal aggressiveness: A review of recent theory and research. *Annals of the International Communication Association* 19, 1 (1996), 319–352.

[43] Dominic A Infante, Bruce L Riddle, Cary L Horvath, and Sherlyn-Ann Tumlin. 1992. Verbal aggressiveness: Messages and reasons. *Communication Quarterly* 40, 2 (1992), 116–126.

[44] KATHERINE ISBISTER and CLIFFORD NASS. 2000. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies* 53, 2 (Aug 2000), 251–267.

[45] Carroll Ellis Izard. 1979. *The Maximally Discriminitive Facial Movements Coding System, MAX*. University of Delaware.

[46] Dr. Rajiv Jhangiani and Dr. Hammond Tarry. 2022. *Principles of Social Psychology – 1st International H5P Edition*. BCcampus, Chapter 9.1.

[47] Joshua Johnson, Sara Hansen, Luke Hopper, Jessica Watson, Sean Cashman, Wyatt De Souza, and Brennen Mills. 2024. Immersive virtual reality aggression and Violence Management Education for Nursing Students: A pre-test, post-test, follow-up evaluation. *Clinical Simulation in Nursing* 97 (Dec 2024), 101644.

[48] Matt Johnston. 2019. Newcastle Uni Taps Biofeedback in VR conflict training. https://www.itnews.com.au/news/newcastle-uni-taps-biofeedback-in-vr-conflict-training-534860

[49] Oswald D. Kothgassner, Andreas Goreis, Lisa M. Glenk, Johanna Xenia Kafka, Bettina Pfeffer, Leon Beutl, Ilse Kryspin-Exner, Helmut Hlavacs, Rupert Palme, and Anna Felnhofer. 2021. Habituation of salivary cortisol and cardiovascular reactivity to a repeated real-life and virtual reality Trier Social Stress Test. *Physiology and Behavior* 242 (Dec 2021), 113618.

[50] Theresa Küntzler, T Tim A Höfling, and Georg W Alpers. 2021. Automatic facial expression recognition in standardized and non-standardized emotional expressions. *Frontiers in psychology* 12 (2021), 627561.

[51] Shu-Min Leong, Raphaël C-W Phan, and Vishnu Monn Baskaran. 2024. Emotion-specific AUs for micro-expression recognition. *Multimedia Tools and Applications* 83, 8 (2024), 22773–22810.

[52] Jill Lobbestael and Maaike J Cima. 2021. Virtual reality for aggression assessment: The development and preliminary results of two virtual reality tasks to assess reactive and proactive aggression in males. *Brain sciences* 11, 12 (2021), 1653.

[53] Elhassan Makled, Christoph Gerhardt, Tobias Schwandt, Florian Weidner, and Wolfgang Broll. 2025. Evaluating behavioral realism in AR and VR: A comparison of single-point IK and full-body motion capture virtual humans. *The Visual Computer* (Apr 2025).

[54] Rachel McDonnell, Martin Breidt, and Heinrich H. Bülthoff. 2012. Render me real? *ACM Transactions on Graphics* 31 (2012), 1–11.

[55] H. Karin M. Meeren, Cindy C. R. J. van Heijnsbergen, and Beatrice de Gelder. 2005. Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences* 102, 45 (2005), 16518–16523.

[56] Albert Mehrabian. 2017. *Nonverbal communication*. Routledge.

[57] Ayelet Melzer, Tal Shafir, and Rachelle Palnick Tsachor. 2019. How do we recognize emotion from movement? Specific motor components contribute to the recognition of each emotion. *Frontiers in psychology* 10 (2019), 392097.

[58] Microsoft Azure. 2022. Azure Custom Neural Voice introduces new emotional styles to enhance customer experience. https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/azure-custom-neural-voice-introduces-new-emotional-styles-to/ba-p/3695831.

[59] Microsoft Research. 2021. Azure AI milestone: New neural text-to-speech models more closely mirror natural speech. https://www.microsoft.com/en-us/research/blog/azure-ai-milestone-new-neural-text-to-speech-models-more-closely-mirror-natural-speech/.

[60] Nathan Moore, Naseem Ahmadpour, Martin Brown, Philip Poronnik, and Jennifer Davids. 2022. Designing virtual reality–based conversational agents to train clinicians in verbal de-escalation skills: Exploratory usability study. *JMIR Serious Games* 10, 3 (2022), e38669.

[61] Christos Mousas, Dimitris Anastasiou, and Ourania Spantidi. 2018. The effects of appearance and motion of virtual characters on emotional reactivity. *Computers*

in Human behavior 86 (2018), 99–108.

[62] Emily Mower, Maja J Mataric, and Shrikanth Narayanan. 2009. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *IEEE Transactions on Multimedia* 11, 5 (2009), 843–855.

[63] Hyeongil Nam, Chanhee Kim, Kangsoo Kim, Ji-Young Yeo, and Jong-Il Park. 2022. An emotionally responsive virtual parent for Pediatric Nursing Education: A framework for multimodal momentary and accumulated interventions. *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Oct 2022), 365–374.

[64] Radosław Niewiadomski, Virginie Demeure, and Catherine Pelachaud. 2010. Warmth, competence, believability and Virtual Agents. *Lecture Notes in Computer Science* (2010), 272–285.

[65] Nahal Norouzi, Kangsoo Kim, Jason Hochreiter, Myungho Lee, Salam Daher, Gerd Bruder, and Greg Welch. 2018. A systematic survey of 15 years of user studies published in the intelligent virtual agents conference. In *Proceedings of the 18th international conference on intelligent virtual agents*. 17–22.

[66] Sung Park and Richard Catrambone. 2021. Social responses to virtual humans: The effect of human-like characteristics. *Applied Sciences* 11, 16 (Aug 2021), 7214.

[67] Catherine Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Communication* 51, 7 (2009), 630–639.

[68] Marc D Pell, Laura Monetta, Silke Paulmann, and Sonja A Kotz. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior* 33 (2009), 107–120.

[69] Marc D Pell, Silke Paulmann, Chinar Dara, Areej Alasseri, and Sonja A Kotz. 2009. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics* 37, 4 (2009), 417–435.

[70] Rosalind W. Picard. 2000. *Affective Computing*. MIT Press.

[71] Inc. Plask. 2024. *Plask.ai: Video to Animation Software*. https://plask.ai/ AI-Powered Animation from Video.

[72] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience* 1 (1980).

[73] Frank E Pollick, Helena M Paterson, Armin Bruderlin, and Anthony J Sanford. 2001. Perceiving affect from arm movement. *Cognition* 82, 2 (2001), B51–B61.

[74] Anat Rafaeli, Amir Erez, Shy Ravid, Rellie Derfler-Rozin, Dorit Efrat Treister, and Ravit Scheyer. 2012. When customers exhibit verbal aggression, employees pay cognitive costs. *Journal of applied psychology* 97, 5 (2012), 931.

[75] Tanmay Randhavane, Aniket Bera, Kyra Kapsaskis, Kurt Gray, and Dinesh Manocha. 2019. Fva: Modeling perceived friendliness of virtual agents using movement characteristics. *IEEE transactions on visualization and computer graphics* 25, 11 (2019), 3135–3145.

[76] Tanmay Randhavane, Aniket Bera, Kyra Kapsaskis, Rahul Sheth, Kurt Gray, and Dinesh Manocha. 2019. Eva: Generating emotional behavior of virtual agents using expressive features of gait and gaze. In *ACM symposium on applied perception 2019*. 1–10.

[77] Reallusion. 2024. *iClone*. https://www.reallusion.com/iclone/ 3D Animation Software.

[78] Ricardo Rodrigues, Ricardo Silva, Ricardo Pereira, and Carlos Martinho. 2022. A cautionary tale of side-by-side evaluations while developing emotional expression for intelligent virtual agents. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery.

[79] Etienne B Roesch, Lucas Tamarit, Lionel Reveret, Didier Grandjean, David Sander, and Klaus R Scherer. 2011. FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *Journal of Nonverbal Behavior* 35 (2011), 1–16.

[80] Astrid Rosenthal-von der Pütten, Julia Arndt, Aleks Pieczykolan, Maria Pohl, and Malte Jung. 2025. Within, between, forced choice, or likert scale? how methodological decisions influence recognition rates in HRI Recognition Studies. *International Journal of Social Robotics* 17, 4 (Mar 2025), 693–706.

[81] Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology* 66, 2 (1994), 310.

[82] Marc Schröder. 2010. The semaine API: Towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Computer Interaction* 2010 (2010), 1–21.

[83] Peggy Schwartz. 1995. Laban movement analysis: Theory and application. *Journal of Physical Education, Recreation & Dance* 66, 2 (1995), 25.

[84] Sinan Sonlu, Uğur Güdükbay, and Funda Durupinar. 2021. A conversational agent framework with multi-modal personality expression. *ACM Transactions on Graphics (TOG)* 40, 1 (2021), 1–16.

[85] Jill E Spielfogel and J Curtis McMillen. 2017. Current use of de-escalation strategies: Similarities and differences in de-escalation across professions. *Social Work in Mental Health* 15, 3 (2017), 232–248.

[86] Donald Stephen and Shahren Ahmad Zaidi Adruce. 2018. Cochran's Q with pairwise mcnemar for dichotomous multiple responses data: A practical approach. *International Journal of Engineering and Technology* 7, 3.18 (Aug 2018), 4.

[87] Bosse Tibor, Gerritsen Charlotte, and de Man Jeroen. 2016. An intelligent system for aggression de-escalation training. *Frontiers in Artificial Intelligence and Applications* (2016).

[88] Christiana Tsiourti, Astrid Weiss, Katarzyna Wac, and Markus Vincze. 2019. Multimodal integration of emotional signals from voice, body, and context: Effects of (in)congruence on emotion recognition and attitudes towards robots. *International Journal of Social Robotics* 11, 4 (Feb 2019), 555–573.

[89] Harald G Wallbott. 1998. Bodily expression of emotion. *European journal of social psychology* 28, 6 (1998), 879–896.

[90] Zijian Wang and Christopher Potts. 2019. TalkDown: A corpus for condescension detection in context. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), 3709–3717.

[91] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. https://arxiv.org/abs/2201.11903

[92] Graham Wilson and Stephen A Brewster. 2017. Multi-moji: Combining thermal, vibrotactile & visual stimuli to expand the affective range of feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1743–1755.

[93] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[94] Xingyu Yang and Kening Zhu. 2023. Emoband: Investigating the affective perception towards on-wrist stroking and squeezing feedback mediated by different textile materials. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–20.

[95] Jiayi Yu. 2023. Research on factors affecting pain perception. *Theoretical and Natural Science* 15 (Dec 2023), 146–151.

[96] Maryam Zahabi and Ashiq Mohammed Abdul Razak. 2020. Adaptive Virtual Reality-based training: A systematic literature review and framework. *Virtual Reality* 24, 4 (Mar 2020), 725–752.

A User Study Details

This section provides supplementary information on both user studies: Experiment 1 (Unimodal Perception) and Experiment 2 (Multimodal Perception).

A.1 Example phrases used in Language modality

Table 8 lists example phrases used for the Language modality across all aggression levels in both the unimodality and multimodality studies.

Table 8: Example phrases representing different levels of aggression

Aggression Level	Example Phrase	Message Types
Not Aggressive	Excuse me, I forgot to order my special meal for this flight. Do you have any extra? Usually, I get the [special menu item] when I travel with [X airline]. I'm a vegetarian and prefer rice or noodles. I'd like to have the [menu item] with a [X drink]. Also, I'll be flying again soon.	None
Slightly Aggressive	Oh, I totally spaced on ordering my special meal for this flight. Any chance you've got an extra one stashed away somewhere? If not, I guess I'll be starting my in-flight fasting routine early. I usually go for the [special menu item]—can't just eat anything, you know? I'm a vegetarian and I need my rice or noodles. But hey, if you're planning to serve me a salad instead, I might just have to share my "delight" in a review online. And just so you know, I don't want this to happen again on my next flight.	Teasing, threats
Aggressive	I want a [special menu item] — think you can manage that, or is that going to fry your last brain cell? It's got to be rice or noodles, no bloody exceptions. And make sure my meal comes with a [X drink]—assuming you even know what the hell that is. Damn it, I wouldn't even be surprised if you have never heard of it, given your background.	Swearing, ridicule, background attacks
Very Aggressive	If you can't get me my vegetarian meal right now, I sure hope you get fired for your incompetence! And it must be [special menu item]. Don't tell me you can't even provide such a simple meal request, did [X airline] even train you properly? I need a vegetarian meal with rice or noodles. Also, make sure it comes with the [X drink]. Judging by how you look, I bet you mess up orders all the time, so don't mess it up!	Maledictions, character attacks, competence attacks, physical appearance attacks

A.2 Survey Questions used for Qualitative Feedback

The open-ended questions used in the multimodality study were as follows:

(1) Which modality (Language, Voice, or Body) was most helpful for judging the IVA's aggression?
(2) How did you feel when different modalities conveyed conflicting aggression levels? Did you experience any difficulty or uncertainty in making your judgment?
(3) When modalities conflicted, did you focus on one modality more than the others, or try to combine information from all cues? How did you resolve the conflict?

A.3 Shapiro–Wilk Normality Tests for statistical measurements

Table 9 and Table 10 shows the Shapiro–Wilk normality test results for all measurements used in the unimodality and multimodality studies respectively.

Table 9: Shapiro–Wilk Normality Test Results for Unimodality measurements

Variable	W	p
Language		
Perceived Aggression	0.817	<0.001
Voice		
Perceived Aggression	0.957	0.011
Loudness	0.967	0.043
Roughness	0.967	0.048
Body Movement & Facial Expressions		
Perceived Aggression	0.946	0.003
Effort-Weight	0.939	0.001
Effort-Time	0.934	<0.001
Effort-Space	0.953	0.007
Shape	0.944	0.002
Lowering Eyebrows	0.956	0.010
Raising Upper Eyelids	0.962	0.024
Tightening Eyelids	0.950	0.005
Tightening Lips	0.976	0.149

Table 10: Shapiro–Wilk Normality Test Results for Multimodality measurements

Variable	W	p
Perceived Aggression	0.986	<0.001