From Multimodal Signals to Adaptive XR Experiences for De-escalation Training

Birgit Nierula[1,*,†], Karam Tomotaki-Dawoud[1,†], Daniel Johannes Meyer[1,†], Iryna Ignatieva[1], Mina Mottahedin[1], Thomas Koch[1] and Sebastian Bosse[1]

[1]*Department for Vision and Imaging Technologies, Fraunhofer Heinrich-Hertz-Institute, Einsteinufer 37, 10587 Berlin, Germany*

Abstract

We present the early-stage design and implementation of a multimodal, real-time communication analysis system intended as a foundational interaction layer for adaptive VR training. The system integrates five parallel processing streams: (1) verbal and prosodic speech analysis, (2) skeletal gesture recognition from multi-view RGB cameras, (3) multimodal affective analysis combining lower-face video with upper-face facial EMG, (4) EEG-based mental state decoding, and (5) physiological arousal estimation from skin conductance, heart activity, and proxemic behavior. All signals are synchronized via Lab Streaming Layer to enable temporally aligned, continuous assessments of users' conscious and unconscious communication cues.

Building on concepts from social semiotics and symbolic interactionism, we introduce an interpretation layer that links low-level signal representations to interactional constructs such as escalation and de-escalation. This layer is informed by domain knowledge from police instructors and lay participants, grounding system responses in realistic conflict scenarios.

We demonstrate the feasibility and limitations of automated cue extraction in an XR-based de-escalation training project for law enforcement, reporting preliminary results for gesture recognition, emotion recognition under HMD occlusion, verbal assessment, mental state decoding, and physiological arousal. Our findings highlight the value of multi-view sensing and multimodal fusion for overcoming occlusion and viewpoint challenges, while underscoring that fusion and feedback must be treated as design problems rather than purely technical ones. The work contributes design resources and empirical insights for shaping human–AI-powered XR training in complex interpersonal settings.

Keywords

Multimodal fusion, Nonverbal cue recognition, Immersive virtual reality, XR-based training, Physiological signal analysis

1. Introduction

Human communication is inherently multimodal, combining verbal content with non-verbal signals such as body posture, gestures, facial expressions, vocal prosody, and interpersonal space [1]. In emotionally charged interactions, these non-verbal cues often carry as much -or more- meaning than spoken language [2]. They are frequently produced unconsciously [3], yet they strongly shape how intentions are perceived and whether an interaction escalates or de-escalates [2]. For professionals operating in high-stakes contexts, such as law enforcement [4] or healthcare [5], the ability to regulate both verbal and non-verbal communication is therefore a critical skill.

Traditional communication training methods [6, 7, 8], including role-play and instructor-led feedback, can support the development of these skills but are limited by subjectivity, delayed feedback, and the difficulty of making unconscious communication patterns visible to trainees. Video-based review and post-hoc analytics offer additional insight, but they separate reflection from the lived, embodied experience of the interaction itself.

Immersive virtual reality (VR) provides a promising medium for communication training by enabling embodied, situated interactions that evoke realistic emotional and behavioural responses while re-

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus

*Corresponding author.

†These authors contributed equally.

Birgit.Nierula@hhi.fraunhofer.de (B. Nierula)

0000-0003-2174-5872 (B. Nierula); 0009-0006-2745-8312 (K. Tomotaki-Dawoud); 0009-0009-6039-7234 (D. J. Meyer)

maining safe, repeatable, and controllable [9]. Prior studies have shown that VR can elicit affective and physiological reactions comparable to real-world encounters [10, 11]. However, designing VR systems for interpersonal skill training introduces a fundamental challenge: communication unfolds continuously across multiple modalities, and its meaning is highly context-dependent [12]. Systems that rely on predefined choices, scripted branching, or post-session evaluation therefore struggle to capture the dynamic nature of real-world interactions.

To shape meaningful training experiences, XR environments must be able to sense and interpret users' communicative behaviour in real time and translate it into adaptive responses that remain immersive and interpretable. This raises a central design question for XR: how can multimodal communication signals be operationalized in a way that supports adaptive, experience-driven training without disrupting the interaction itself? Addressing this question requires not only accurate signal processing, but also careful design decisions regarding which cues matter, how they are interpreted, and how they influence virtual character behaviour and feedback.

Research Gap. Limitations of current solutions are their lack of objective feedback and scalability as they are typically implemented as role-play and instructor-led feedback. Objective and scalable XR solutions that allow to capture and predict the dynamics of social interactions are challenged by a lack of multimodal (verbal and non-verbal) communication sensing in real-time in a context-aware manner. This, however, is necessary for adaptive, context-aware avatar responses that combined with objective and transparent feedback can foster awareness and control over own verbal and non-verbal communication cues. Implementing this requires both technical signal processing capabilities and social theory design frameworks for operationalizing communication cues.

Contributions. This work offers the following contributions to the design of multimodal communication sensing XR systems grounded in social science:

- **Implemented and evaluated:**
  - A multi-view skeletal gesture recognition pipeline achieving 82–88% accuracy for conflict-relevant postures and gestures;
  - A late-fusion architecture combining lower-face video with upper-face EMG for emotion recognition under HMD occlusion, achieving 51% macro-F1 on 7-class classification;
  - Physiological arousal analysis (SCR, HRV) validated in a controlled experimental scenario with sensitivity to proxemic violations and emotional facial expressions;
  - Verbal communication analysis including formality detection (68% accuracy)
  - prosodic feature extraction with preliminary correlation analysis.
- **Implemented with preliminary evaluation:**
  - EEG-based mental state decoding using SSD+SPoC decomposition, demonstrated as proof-of-concept in 3 participants;
  - Sentence complexity assessment, evaluated on external benchmark data but not yet validated in police interaction contexts.
- **Conceptual contributions:**
  - An integrated system architecture synchronizing five parallel analysis streams via Lab Streaming Layer for temporally aligned multimodal assessment;
  - A design framework grounded in social semiotics and symbolic interactionism, linking low-level signal representations to interactional constructs (escalation/de-escalation) through domain expert and lay participant input;
  - Design considerations for multimodal fusion and adaptive feedback in XR training, framed as open design problems rather than solved technical challenges.

By distinguishing evaluated components from preliminary and conceptual contributions, we aim to provide transparency about the current maturity of each system element.

We illustrate this approach through its application in an ongoing research project on de-escalation training for law enforcement. We report preliminary findings and design reflections that suggest the feasibility of using multimodal communication analysis to inform real-time adaptation in immersive environments. By foregrounding design considerations and early insights, this work contributes to the ShapeXR community by highlighting how human–AI-powered XR experiences can be used to support embodied learning in complex interpersonal scenarios.

2. Related Work

In recent years, a range of product developments and research activities have sought to exploit the advantages of XR-based training in the policing context. The TARGET innovation project [13] developed a toolkit for creating integrated AR/VR training scenarios, which also enables the representation of complex, pan-European large-scale operations. Similarly, many existing training solutions focus on scenario simulation to train procedures, behavioural patterns, decision-making, and communication from a tactical, process-oriented perspective [14, 15, 16, 17]. These approaches often incorporate haptic elements through the use of operational equipment, particularly weapons [18]. Data-driven tactical and behavioural analyses are typically conducted after the training, based on video data, for example [19]. In contrast to these existing trainings, our approach does not aim to model the operational process as a whole, but instead focuses on communication dynamics in critical situations with the aim to avoid the use of operational equipment, particularly weapons.

In the domain of communication training, both commercial products and research contributions demonstrate the benefits of data-driven communication analysis. For instance, VR trainings for presentation training incorporate speech-related quality metrics to provide automated post-training evaluations [20]. In the context of police training, research shows the usefulness of VR communication training for critical situations, as well as the benefit of and demand for evaluation and visualization of the training [21] [22]. As an interactional feedback mechanic, in 2013 a job interview simulator [23] leveraged the recognition of nonverbal communication cues [24] not only to generate useful feedback but also to model specific communication dynamics within training scenarios in real-time. While existing studies have established the potential value of such approaches in principle [25], interactive feedback typically focuses on specific aspects of communication. To allow progress towards a more generalizable model of communication, the presented approach develops a system in which a broad range of relevant modalities, combined with potentially interpretative signal analysis, serves as the foundation for designing a more comprehensive feedback system for VR trainings.

Recent work in applied virtual environments has investigated combining neural and peripheral physiological signals for user state assessment, including simulator sickness [26] and perceived face realness [27]. Real-time multimodal analysis enables affect-adaptive systems; for instance, EEG and peripheral signals have been used to modulate game difficulty [28]. In parallel, facial electromyography (EMG) has emerged as a robust modality under HMD occlusion: wearable platforms such as EmteqPRO integrate upper-face EMG with physiological sensors in VR-compatible form factors [29, 30], enabling expression classification [31] and continuous valence-arousal monitoring during immersive viewing [32]. Complementary optical approaches address occlusion through peri-ocular IR imaging [33], photo-reflective sensors [34], or generative face completion [35]. Beyond facial cues, recent work highlights the role of body posture and gesture as salient non-verbal communication markers in interactive and conflict-related scenarios [36]. Multimodal fusion strategies have shown consistent improvements over unimodal baselines in VR-based emotion, action, and user-state recognition [37, 38].

3. Extraction Layer

We propose an integrated pipeline that extracts communication cues from multimodal sensor data to enable comprehensive feedback in immersive VR training environments. The system operates on five parallel processing streams: (1) a speech analysis stream including both verbal and prosodic features, (2) a skeletal gesture stream capturing body posture and movement dynamics [36], (3) a multimodal affective stream integrating facial expressions and physiological signals [39], (4) mental state decoding stream and (5) a bodily arousal stream. This architecture enables real-time assessment of both conscious and unconscious communication cues, which can be essential for professional training in domains such as law enforcement de-escalation and healthcare communication.

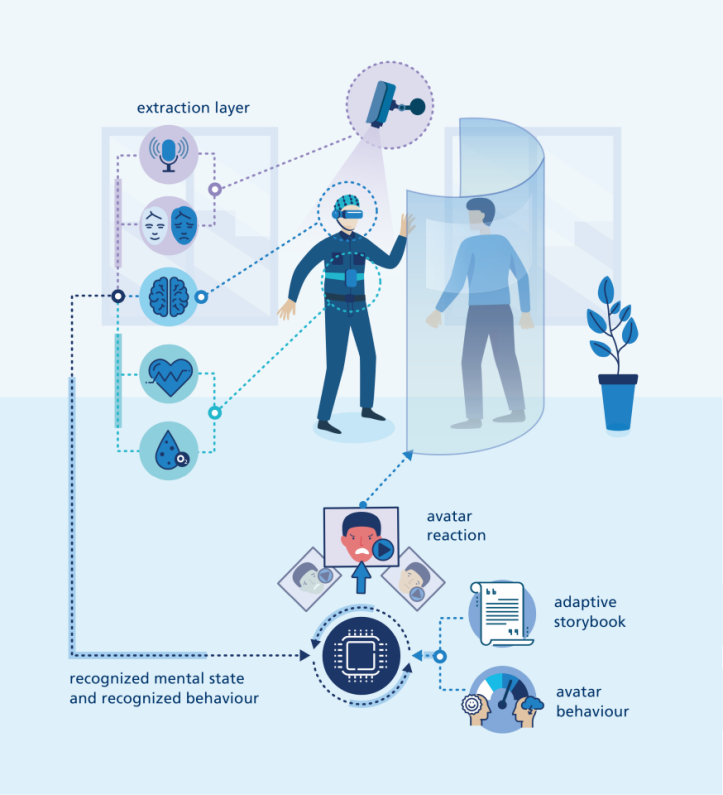

Figure 1: Conceptual system overview of the extraction layer integrated into an adaptive XR experience. The system objectively assesses user experiences during virtual conflicts by extracting features from signal modalities and fusing them with avatar behaviour and storybook context. The resulting signal adapts the virtual experience.

3.1. System Overview

The pipeline begins with synchronized acquisition from multiple sensors: (i) at least one external microphone, (ii) two or more cameras positioned at complementary angles (e.g., 0° and ±30° relative to the sagittal plane) for capturing body gestures and lower-face expressions, and (iii) a VR-integrated head-mounted display (HMD) equipped with facial electromyography (EMG) sensors on the upper face, (iv) electroencephalography, (v) skin conductance, and (vi) electrocardiography (see Figure 1 for system overview).

All data streams are synchronized via Lab Streaming Layer (LSL [40, 41]) with event markers inserted at critical training milestones, ensuring precise temporal alignment.

3.2. Analysis Streams

Verbal Communication The verbal content of the audio signal is extracted by faster-whisper [42], an optimized reimplementation of OpenAI's Whisper model [43]. Insults are detected by keyword recognition. For complex classification tasks, a transformer-based language model is employed. Our first application focuses on formality detection, since, for example, using the informal "du" instead of "Sie" may be perceived as disrespectful. For this, we used a transformer-based language model (XLM-R-large). The model was fine-tuned on a subset of German data from the FAME-MT [44] and CoCoA-MT [45] datasets to classify user utterances as formal, informal, or neutral.

Given its relevance in the context of language barriers, we explored sentence complexity assessment. For this subtask we followed the approach of Asghari and Hewett (2022) [46] and use both linguistic features and a XLM-R-large model fine-tuned on Text Complexity DE [47] dataset to predict how complex a sentence is on a scale from 1 (least complex) to 7 (most complex).

Nonverbal Speech Parameters Closely connected to our perception, prosodic features play an important role in the impact of spoken language. From the microphone signal, a fixed calibration extracts the absolute sound pressure at a reference distance of 1 meter, allowing for the calculation of psychoacoustic features. Highly connected to the affective content of speech are loudness and pitch [48]. Loudness is calculated according to ISO-standard [49] using the MOSQITO-library [50]. The PYIN algorithm is used [51] to estimate a pitch contour, from which the average pitch as well as minimum and maximum pitch are derived. To extract valence and arousal a speech emotion recognition model is applied [52]. Based on the transcription of the faster-whisper model, the speaking rate is calculated.

Skeletal Gesture Recognition RGB images from multiple camera views are processed through BlazePose [53, 54, 55] for real-time 3D pose estimation, extracting body landmarks that serve as kinematic skeletons. Features are derived from pairwise Euclidean distances between 3D joint coordinates, normalized to account for individual anthropometric variation. This representation captures the spatial relationships underlying conflict-relevant gestures (e.g., crossed arms) without dependency on appearance, where simple neural networks classify these normalized distance features into discrete gesture categories, enabling immediate recognition of escalation or de-escalation signals (according to context described in Sec. 4).

Emotional Facial Expressions This stream operates on two complementary modalities addressing the HMD occlusion challenge. The visual branch extracts lower-face features via ResNet-18/50 backbones pre-trained on CK+ with occlusion augmentation (black patches, random noise, VR headset overlays) to focus on mouth and cheek regions. The physiological branch processes seven-channel facial EMG (corrugator supercilii, frontalis, orbicularis oculi, zygomaticus major) using a compact RBF kernel representation that captures inter-muscle correlations within sliding windows. These correlations reflect the complex multi-muscle patterns underlying emotional expressions.

Visual and EMG embeddings are independently projected to 128-D representations and then concatenated into a 256-D fused feature vector, which is passed through fully connected layers to classify seven emotion categories (the six basic emotions introduced by Ekman [56] plus neutral). This modular late-fusion design enables independent optimization of each stream while allowing seamless incorporation of additional modalities (e.g., skin conductance, heart rate) without architectural changes.

The weighted combination of gesture and emotion recognition streams (shown in fig. 2) enables VR training systems to provide real-time feedback, where gesture recognition raises trainee awareness of unconsciously performed body language, while emotion recognition reflects the emotional impact of their communication, facilitating iterative skill refinement in realistic simulated scenarios.

Mental State Decoding A user's mental state can be extracted via electroencephalography (EEG), a non-invasive technique that measures cortical electrical activity through scalp sensors. To isolate relevant neural signals from the complex mixture of brain activity, we employed multivariate source decomposition methods. These computational approaches enable extraction of specific cortical activity patterns from 64-channel EEG recordings in relation to arousal-inducing stimuli. As a first target variable, we used the avatar's proxemic behavior—specifically, the avatar approaching the user.

Specifically, we employed a two-stage decomposition pipeline combining spatio-spectral decomposition (SSD [57]) and source power comodulation (SPoC [58]) to extract arousal-related cortical activity induced by avatar approach. In the first stage, SSD was applied to maximize the signal-to-noise ratio in the alpha frequency band (8-12 Hz), which is known to inversely correlate with arousal states in frontal and parieto-occipital cortical regions. SSD identifies spatial filters that enhance oscillatory components within this target frequency range while simultaneously suppressing activity from flanking frequency bands, thereby isolating spectrally pure alpha oscillations from the multi-channel EEG data. In the second stage, SPoC was applied to the SSD-filtered components to identify neural sources whose alpha power fluctuations showed maximal covariation with the time-varying distance of the approaching avatar. SPoC optimizes spatial filters by finding linear combinations of EEG channels that maximize the

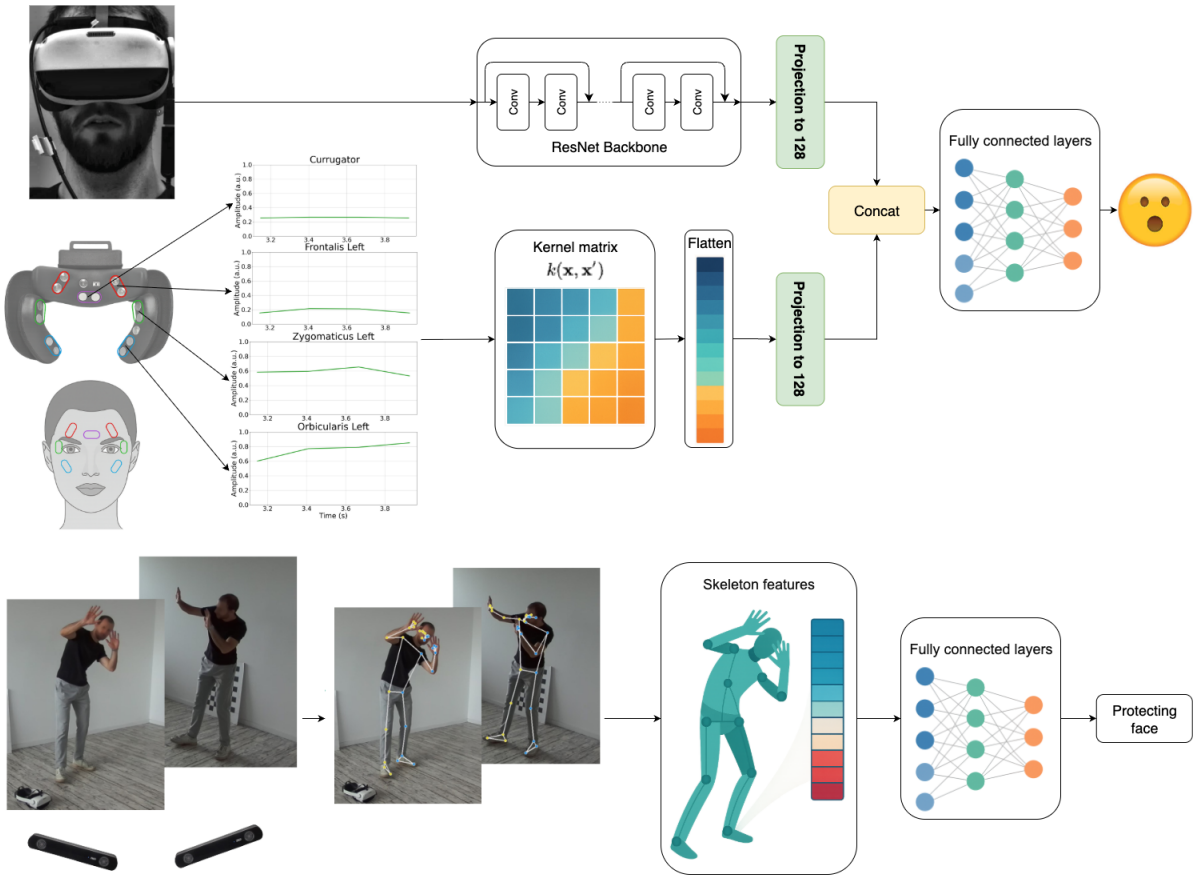

Figure 2: Integrated dual-stream pipeline combining gesture and emotion recognition for VR training feedback. The emotion stream (top) fuses lower-face video with upper-face EMG under HMD occlusion; the gesture stream (bottom) processes multi-view skeletal data to detect conflict-relevant body language. Both streams converge into a unified weighted output that delivers real-time, multimodal feedback for skill refinement.

correlation between extracted source power and the continuous target variable (avatar proximity). This approach has been shown to identify arousal-sensitive neural patterns in VR users [59].

Bodily Arousal Recognition Emotionally charged interactions trigger physiological arousal that can be objectively measured through skin conductance response (SCR) and heart activity. These two continuous signals can capture distinct aspects of the autonomic nervous system response during emotionally charged interactions. SCR reflects sympathetic activation, with higher normalized peak amplitudes indicating greater arousal, while heart rate variability (HRV) reflects parasympathetic activity, with lower values indicating autonomic stress responses. Our analysis pipeline returns (1) an SCR stream providing normalized peak amplitudes and (2) a HRV stream providing the Root Mean Square of Successive Differences (RMSSD) of R-R intervals. The pipeline is described in more detail in [60].Individual baseline measurements recorded prior to the interaction establish participant-specific thresholds to detect when physiological signals deviate from resting state, indicating arousal.

Proxemic Regulation Proxemic behaviour, the spatial positioning between user and avatar, is extracted from the VR headset and avatar positions. The interpersonal distance reflects how individuals regulate space during emotional interactions and serves as a non-verbal indicator of comfort, threat perception, and social engagement.

3.3. Illustrative Example

To clarify how the extraction layer informs adaptive behavior, consider the following hypothetical scenario during a de-escalation training session:

A trainee officer attempts to calm a visibly agitated virtual citizen. As the interaction progresses, the system continuously monitors multiple streams (see Table 1).

Table 1
Example of multimodal cues observed during a hypothetical training interaction.

Stream	Observed Cue	Extracted Feature
Verbal	Officer uses informal address ("du")	Formality classifier → *informal*
Prosody	Elevated vocal loudness (mean loudness > 8 Sone)	Loudness module → *high*
Gesture	Arms crossed, weight shifted backward	Pose classifier → *defensive posture*
Physiology	SCR amplitude 1.5× baseline	Arousal module → *elevated*

These signals are forwarded to the interpretation layer, where context from the scenario storybook (e.g., "citizen is non-native speaker; prior exchange was tense") modulates their meaning. The co-occurrence of informal language, elevated loudness, and defensive posture (combined with rising physiological arousal) is interpreted as emerging trainee frustration with potential escalation risk.

Based on this interpretation, the system triggers a conservative avatar response: the virtual citizen takes a small step backward, averts gaze briefly, and lowers vocal intensity; subtle behavioral shifts designed to signal discomfort without abrupt scene changes. A cooldown window (e.g., 5 seconds) prevents repeated adaptations from transient signal fluctuations. Post-scenario, the trainer reviews flagged moments alongside the multimodal timeline, contextualizing feedback for the trainee.

This example illustrates the intended information flow; however, as discussed in Section 6, the fusion logic and adaptation policies remain under active development.

3.4. Method Summary

Table 2 provides a compact overview of the core processing parameters for each analysis stream. The extraction layer follows a modular design philosophy that prioritizes methodological transparency while maintaining deployment flexibility.

We intentionally refrain from providing a fixed instrumentation specification (e.g., exact device models, proprietary sensor configurations) because the pipeline is designed to be hardware-agnostic within defined functional constraints. Specific sensors—including cameras, microphones, EEG systems, and physiological recording platforms—can be exchanged depending on deployment context, facility requirements, and evolving technology. For instance, any RGB camera system meeting minimum frame-rate requirements ($\geq$30 fps) can serve the gesture stream, and various commercial EEG systems supporting 32+ channels and LSL integration are compatible with the mental state decoding pipeline. Similarly, facial EMG can be acquired through standalone electrode arrays or integrated HMD-mounted platforms.

This flexibility is a deliberate design choice: by decoupling the signal-processing methodology from specific hardware implementations, the system remains applicable across heterogeneous training environments—from laboratory settings to operational training facilities—and can incorporate improved sensors as they become available. The LSL-based synchronization architecture ensures temporal alignment regardless of the specific device combination, provided each stream publishes timestamped data to the shared network.

Minimum functional requirements for each modality are as follows: audio input at $\geq$48 kHz sampling rate, capable of calibration (e.g. without integrated auto-leveling or noise reduction); RGB video at $\geq$30 fps with sufficient resolution for pose estimation ($\geq$1080p); physiological signals (ECG, EDA) at $\geq$250 Hz; facial EMG at $\geq$800 Hz; and EEG at $\geq$250 Hz with a minimum of 32 channels for source decomposition methods.

Table 2
Compact method summary for each analysis stream. Specific hardware can be exchanged within stated functional constraints; see text for minimum requirements.

Stream	Preprocessing / Artifact Management	Window / Segmentation	Feature Extraction
Verbal	High pass (60 Hz); speech-to-text (faster-whisper)	Sentence-wise (based on STT output)	Transformer embeddings (XLM-R-large); keyword matching for insult detection; formality and complexity scores
Prosody	High pass (60 Hz); fixed calibration for absolute sound pressure at reference distance of 1 meter; voice activity detection. *For speaking rate:* speech-to-text	Continuous sliding window (e.g. 2.0 s window and 0.5 s hop). *For speaking rate:* sentence-wise separation	Loudness (mean of time-varying loudness according to ISO 532-1, via MOSQITO); pitch contour (PYIN: mean, min, max); valence/arousal (wav2vec2-based SER); speaking rate (wpm)
Gesture	Multi-view frames synchronization and normalization	Per-frame (~33 ms at 30 fps)	33 BlazePose landmarks → pairwise 3D Euclidean distances; normalized feature vector input to Random Forest classifier
Facial Emotion (Multimodal)	*EMG:* Notch filter (50 Hz + harmonics), bandpass 100-400 Hz; baseline subtraction; amplitude clipping; full-wave rectification; min–max scaling. *Video:* Face detection and cropping; intensity normalization.	*EMG:* 1.0 s window, 0.25 s hop. *Video:* Single frame per window (temporally aligned with EMG)	*EMG:* 7-channel inter-muscle RBF kernel matrix → 128-D projection. *Video:* ResNet-50 → 128-D projection. *Fusion:* Late fusion; 256-D concatenated embedding → FC layers → 7-class logits
Mental State (EEG)	Bandpass 1–40 Hz; bad-channel interpolation; ICA-based ocular and muscle artifact removal	1 second, 50% overlap; continuous for real-time	SSD → alpha-band (individual peak *pm*2 Hz) enhancement; SPoC → components maximally correlated with target variable (avatar distance); alpha power as output feature
Bodily Arousal (EDA)	Band-pass filter (0.0159–0.5 Hz); z-score normalization	Event-locked windows (1–5 s post-stimulus); continuous phasic monitoring	Phasic SCR amplitude
Bodily Arousal (ECG)	z-score transformation; high-pass filter $\geq$ 0.5 Hz; R-peak detection; ectopic and missing peak correction	5 seconds; no overlap	RMSSD (root mean square of successive R-R differences)
Proxemics	Coordinate transformation to common reference frame	Continuous (real-time, per VR frame)	Euclidean distance between HMD position and avatar centroid; rate of change (approach velocity)

4. From Semantics to Interpretation

To develop an adaptive system that captures face-to-face interaction in VR, this work draws on insights from social semiotics and symbolic interactionism. Social semiotics provides a framework for decomposing interaction into its multimodal modes and understanding how meaning emerges from their combination [61]. However, meaning is not solely encoded in semantic units but is also con-

structed through interactional processes. Symbolic interactionism posits that meaning is produced and continuously reshaped through participants' mutual interpretation during interaction [62]. From this perspective, interactions are inherently co-constructed: participants continuously interpret and adjust their behavior in response to one another's communicative acts.

The design approach pairs detailed semantic analysis of each modality (from social semiotics) with an interactionist perspective that frames escalation and de-escalation as emergent system dynamics. First, modality elements meaningful for system functionality are identified, translating semantic categories into computational representations. Second, the interactional contexts in which these elements occur are captured. Third, the system examines whether these elements are interpreted differently depending on context—reflecting the symbolic interactionist premise that meaning emerges through situated interpretation rather than fixed semantic codes.

This approach is illustrated through a case study on non-verbal cues. Drawing from symbolic interactionism, escalation or de-escalation in police-citizen encounters are often shaped by turning points marked by shifts in nonverbal behavior [63, 64]. Based on literature and interviews with police trainers, 19 gestures were identified and categorized by communicative function (routine/alert, calming, commanding, space-controlling, reactive). A machine learning pipeline enables real-time detection and classification of these gestures (Section 3.2). Ten realistic scenarios in which these gestures occur were identified through trainer interviews and provide the context information. An online study with German residents (N=600, stratified by age and gender) then assessed how each gesture could escalate or de-escalate interactions within each context.

This process generates escalation/de-escalation indices for each gesture-scenario pairing, differentiated by officer gender and citizen demographics. These theory-driven mappings operationalize the research framework in three stages: (1) multimodal cues (in the case study gestures) are computationally detected, (2) context-dependent interpretive states (escalation/de-escalation potential) are derived from empirical data reflecting situated meaning-making, and(3) adaptive system responses are triggered—enabling virtual agents to escalate or de-escalate based on trainee behavior, adjusting narrative dynamics according to both character and trainee demographics. The indices thus serve as computational parameters that translate interactionist principles into concrete system adaptations.

5. Results

Our investigations into automated recognition of communication cues for conflict-related training scenarios yielded promising results across six complementary domains:

For **verbal communication**, pilot test data was collected from N=20 police officers interacting in a prototypical VR scene for de-escalation training. Formality classification on the mentioned test set achieved an accuracy of 0.68 and a weighted F1-score of 0.64, indicating limited robustness in capturing formality cues beyond simple lexical patterns. This suggests that the model relies heavily on patterns from the training data and leaves room for improvement through more diverse examples and deeper stylistic features.

For **nonverbal speech parameters**, a preliminary exploratory data analysis on the pilot test data was performed. Figure 3 shows the mean length and loudness over the answers of each of the 20 police officers. A correlation of $r = 0.46$ was found, suggesting a dependency of these variables. Given that each participant was pro-

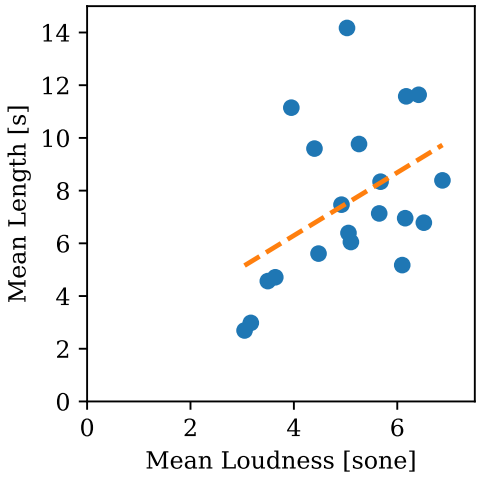

Figure 3: Mean values of length and loudness over answers of N=20 participants with an indicated linear regression.

fessionally trained, there still seems to be significant variation. As a combined measure it might be connected to different communication approaches, as for example shorter, quieter answers might hint at a more cautious approach, while louder, longer answers might hint at a more dominant communication.

For **body gesture recognition**, the multi-view skeleton-based system demonstrated robust performance, with Random Forest classifiers consistently achieving 82–88% accuracy across different feature sets in 5-fold cross-validation. Handcrafted features derived from pairwise 3D joint distances performed comparably to models trained on raw skeleton coordinates, while offering improved interpretability and reduced overfitting. Notably, integrating multi-view camera data enhanced recognition accuracy, particularly in challenging scenarios involving occlusions or non-frontal poses.

For **facial emotion recognition** under HMD occlusion, conventional image-based methods (e.g., OpenFace3 [65]) collapsed to near-chance performance (≈ 16% accuracy on a 7-class prediction) when applied to lower-face-only inputs, highlighting the limitations of transferring existing facial expression recognition pipelines into immersive VR environments. However, our proposed late-fusion architecture—combining lower-face video with seven-channel upper-face EMG—substantially outperformed unimodal baselines, achieving 51% macro-F1 on subject-independent held-out test identities. Unimodal EMG alone achieved 43–46% F1, demonstrating the discriminative strength of facial EMG as a complementary modality. Highly expressive emotions such as happiness and surprise were recognized most reliably (up to 0.84 and 0.75 true positive rates, respectively), whereas subtler expressions (anger, disgust) remained more challenging due to overlapping muscle activation patterns.

For **mental state decoding**, we present preliminary results from 3 participants. Figure 4A shows the SSD-derived spatial filters, which successfully isolated alpha-band oscillations (8-12 Hz). Subsequent SPoC analysis (Figure 4B) identified neural sources whose alpha power fluctuations showed maximal covariation with changes in avatar distance. All participants exhibited SPoC patterns with prominent parieto-occipital activity, consistent with neural correlates of emotional arousal [59]. However, the ranking of SPoC components by correlation strength varied across participants. Understanding the optimal selection strategy of SPoC patterns is subject of ongoing analysis. The time course panel (Figure 4C) illustrates decoding performance for one representative participant (sub-096): the predicted avatar distance reconstructed from SPoC-extracted alpha power (blue) tracks the actual avatar distance (red) with r=0.49.

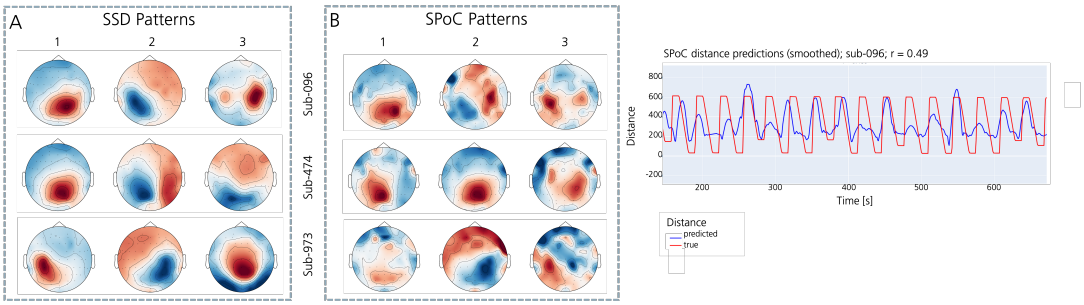

Figure 4: Activation patterns in 3 participants related to emotional arousal adapted from [66].

For **bodily arousal**, physiological markers of emotional arousal were tested in an experimental scenario where the avatar did or did not respect the user's personal space and showed a neutral or angry emotional facial expression [60]. Multimodal assessment revealed that our analysis methods were sufficiently sensitive to detect distinct autonomic nervous system responses across modalities. Skin conductance response (SCR) analysis proved sensitive to capture subtle changes in sympathetic arousal due to personal space violations (see Figure 5B and C), while remaining unaffected by facial expressions. Conversely, heart rate variability (HRV) analysis demonstrated sufficient sensitivity to detect parasympathetic modulation in response to threatening facial cues, but showed no response to spatial violations (see Figure 5D).

Eye tracking evidence further demonstrates the experiential relevance of proxemic signals: interpersonal distance to a virtual instructor significantly affected visual attention patterns, with greater

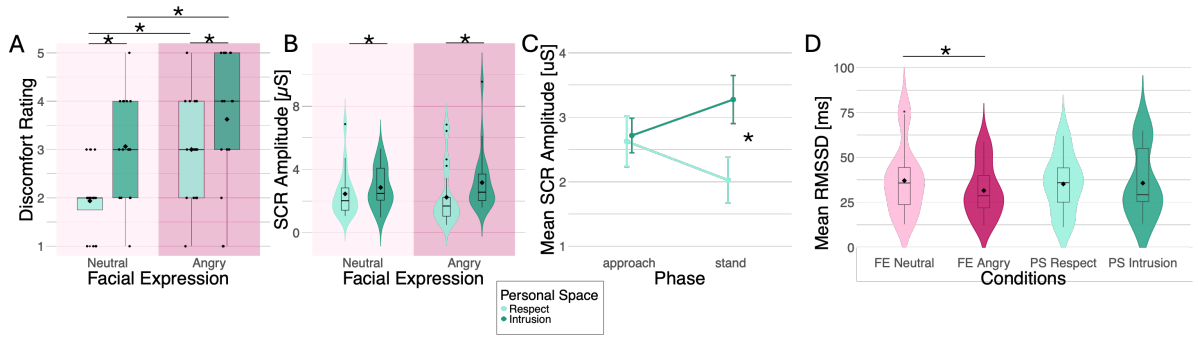

Figure 5: Physiological results adapted from [60]. (A) Discomfort ratings to an avatar that respected the user's personal space (light green) or violated it (dark green) with neutral or angry facial expression. (B) Effect on SCR amplitudes. There was only a main effect of Personal Space on SCR amplitudes. (C) Interaction effect of the avatar phase (whether approaching or standing in front of the user) on SCR amplitudes. (D) Effect on heart rate variability (HRV, measured via RMSSD). Asterisks in all panels indicate significant statistical effects ($p<.05$).

distance leading to attentional dispersion and reduced learning performance [67].

Collectively, these results establish the feasibility of automated non-verbal cue recognition for XR-based training applications and underscore the value of multi-view sensing and multimodal fusion strategies for overcoming occlusion and viewpoint challenges inherent to immersive environments.

6. Discussion

This work presents an early-stage exploration of how multimodal, real-time communication analysis can be used to shape adaptive XR training experiences for de-escalation. Rather than framing multimodal sensing as an end in itself, we treat it as an interaction layer whose interpretation directly influences avatar behaviour and the trainee's lived experience. Our discussion therefore focuses on design implications that arise when transforming heterogeneous communication signals into meaningful, experience-driven adaptations in XR.

A key insight is that multimodal signals only become meaningful through interpretation within a specific interactional context. Gestures, prosody, physiological arousal, and spatial behaviour do not map directly to communicative intent, but must be translated into experiential states such as perceived threat, calmness, or authority. In XR training, these mappings shape how virtual characters respond and whether interactions feel plausible and pedagogically useful. As a result, decisions about which cues matter and how they influence adaptation are fundamentally design choices, not purely technical ones.

Multimodal fusion emerged as a central and unresolved design challenge. The modalities considered in this work operate on different temporal scales, vary in reliability, and may produce conflicting signals. Static or globally optimized fusion strategies risk overreacting to transient signals or privileging modalities that are experientially misleading in a given context. As our fusion approach is still conceptual, we intentionally frame fusion as a design space rather than a finalized solution. Context-sensitive weighting, temporal smoothing, and conservative adaptation strategies appear particularly important for maintaining plausibility and user trust in immersive training scenarios. From an XR perspective, fusion should support experience shaping rather than maximize classification confidence. Importantly, uncertainty in multimodal fusion is not only a technical limitation but also an ethical concern, as overconfident interpretations risk attributing intent, aggression, or communicative failure where none exists.

In the current implementation, each analysis stream operates independently, producing parallel feature outputs that are synchronized via LSL timestamps. While the architecture is designed to support multimodal fusion—where correlations, temporal dependencies, and potential conflicts between

streams would inform unified interpretations—this integration remains conceptual at the present stage. Specifically, we have not yet implemented mechanisms for (1) weighting modalities based on context-dependent reliability, (2) resolving contradictory signals (e.g., relaxed posture but elevated physiological arousal), or (3) learning cross-modal temporal patterns that may be more predictive than any single stream.

We intentionally frame fusion as an open design space rather than a solved problem. Future work will explore both rule-based and data-driven fusion strategies, investigate how scenario context (encoded in the adaptive storybook) can dynamically adjust modality weights, and evaluate how different fusion approaches affect trainee experience and pedagogical outcomes. Until such strategies are validated, the system outputs per-stream features for human review rather than fully automated adaptation decisions.

Another important consideration concerns feedback and immersion. While real-time adaptation of avatar behaviour can provide powerful implicit feedback, overly frequent or explicit responses risk disrupting emotionally charged interactions. Our approach therefore favours embedding feedback primarily in avatar behaviour, complemented by post-scenario reflection and instructor-mediated interpretation. Deciding when not to adapt is as critical as deciding how to adapt, underscoring the importance of restraint and pacing in adaptive XR design. Interpreting emotional, physiological, and behavioural signals in real time raises ethical challenges that directly shape system design. Physiological arousal or non-verbal behaviour cannot be unambiguously mapped to intent or escalation risk, and such signals are sensitive to individual, cultural, and situational factors. To mitigate the risk of over-interpretation, we intentionally favour conservative, context-sensitive adaptations and avoid treating system outputs as objective assessments. The system is designed to support reflection rather than evaluation, with trainers retaining responsibility for contextualizing feedback. This human-in-the-loop approach reflects both ethical considerations and the inherently ambiguous nature of interpersonal communication.

Given the sensitivity of de-escalation training, full automation is neither desirable nor appropriate. We position our system as decision-support rather than decision-making, with human oversight remaining central. Trainers contextualize feedback, and scenario authors define interpretation and fusion rules aligned with pedagogical goals. This highlights the need for XR authoring tools that make multimodal interpretation transparent and adjustable, enabling designers to shape system behaviour without relying on opaque black-box models. Ethical considerations such as privacy, consent, and the risk of misclassification further reinforce the importance of keeping trainees and instructors in control and framing feedback as formative rather than evaluative.

This work reflects an ongoing research effort with several limitations. The presented findings are preliminary and are primarily intended to inform design decisions rather than providing a comprehensive evaluation of training effectiveness or measuring training impact. Future work will focus on refining multimodal fusion strategies, exploring adaptive and personalized interpretations, and conducting larger-scale and longitudinal studies. While our case study centres on law enforcement, the design considerations discussed here are applicable to a wide range of XR training scenarios involving complex interpersonal interaction, including healthcare, education, and negotiation.

Acknowledgments

We would like to express our deepest gratitude to A. Bock, M. Heinze, D. Knuth, D. Martin, J. Schander, A. Orinsky, L. Braunisch, M. Ernst, D. Guthor, J. Gutman and our associated project partners Bayerische Polizei and Polizei Berlin for their valuable insight and guidance. This research was funded by the German Federal Ministry of Education and Research under the grants K3VR (13N16388).

Declaration on Generative AI

During the preparation of this work, the authors used different commercial LLMs for grammar and spelling checking. After using these tools, the authors reviewed and edited the content as needed and

take full responsibility for the publication's content.

References

[1] C. Jewitt (Ed.), The Routledge Handbook of Multimodal Analysis, 2 ed., Routledge, London, 2014.

[2] N. Otu, Decoding nonverbal communication in law enforcement, Salus Journal 3 (2023) 1–16. URL: https://journals.csu.domains/index.php/salusjournal/article/view/42.

[3] J. L. Lakin, Automatic cognitive processes and nonverbal communication, in: V. Manusov, M. L. Patterson (Eds.), The SAGE Handbook of Nonverbal Communication, SAGE Publications, Inc., 2006, pp. 59–78. doi:10.4135/9781412976152.n4.

[4] R. M. Rozelle, J. C. Baxter, The interpretation of nonverbal behavior in a role-defined interaction sequence: The police–citizen encounter, Journal of Nonverbal Behavior 2 (1978) 167–180. doi:10.1007/BF01145819.

[5] M. Schmid Mast, G. Cousin, The role of nonverbal communication in medical interactions: Empirical results, theoretical bases, and methodological issues, in: L. R. Martin, M. R. DiMatteo (Eds.), The Oxford Handbook of Health Communication, Behavior Change, and Treatment Adherence, Oxford Library of Psychology, Oxford University Press, Oxford, 2013. doi:10.1093/oxfordhb/9780199795833.013.021.

[6] M. D. White, C. Orosco, S. Watts, Beyond force and injuries: Examining alternative (and important) outcomes for police De-escalation training, Journal of Criminal Justice 89 (2023) 102129. doi:10.1016/j.jcrimjus.2023.102129.

[7] D. Sjöberg, Simulation Exercises in Police Education, Why and How? A Teacher's Perspective, International Journal for Research in Vocational Education and Training 11 (2024) 460–482. doi:10.13152/IJRVET.11.3.6.

[8] A. Gelis, S. Cervello, R. Rey, G. Llorca, P. Lambert, N. Franck, A. Dupeyron, M. Delpont, B. Rolland, Peer Role-Play for Training Communication Skills in Medical Students: A Systematic Review, Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare 15 (2020) 106–111. doi:10.1097/SIH.0000000000000412.

[9] M. V. Sanchez-Vives, M. Slater, From presence to consciousness through virtual reality, Nature reviews neuroscience 6 (2005) 332–339.

[10] R. B. H. Tootell, S. L. Zapetis, B. Babadi, Z. Nasiriavanaki, D. E. Hughes, K. Mueser, M. Otto, E. Pace-Schott, D. J. Holt, Psychological and physiological evidence for an initial 'Rough Sketch' calculation of personal space, Scientific Reports 11 (2021) 20960. doi:10.1038/s41598-021-99578-1.

[11] J. Marín-Morales, J. L. Higuera-Trujillo, J. Guixeres, C. Llinares, M. Alcañiz, G. Valenza, Heart rate variability analysis for the assessment of immersive emotional arousal using virtual reality: Comparing real and virtual scenarios, PLOS ONE 16 (2021) e0254098. doi:10.1371/journal.pone.0254098.

[12] R. Hodge, G. Kress, Social Semiotics, Polity Press, Oxford, 1998.

[13] TARGET – Training Augmented Reality Generalised Environment Toolkit, https://researchportal.list.lu/projects/detail/target/, 2025. Project webpage. Accessed: 2025-12-19.

[14] RE-liON, Virtual reality training platform, https://re-lion.com, 2025. Company website. Accessed: 2025-12-19.

[15] Refense, Virtual reality training platform, https://www.refense.com, 2025. Company website. Accessed: 2025-12-19.

[16] Street Smarts VR, Virtual Reality Training, https://www.streetsmartsvr.com/vrtraining, 2025. Company website. Accessed: 2025-12-19.

[17] Axon Enterprise, Inc., Axon | Protect Life, https://www.axon.com, 2025. Company website. Accessed: 2025-12-19.

[18] Operator XR, Operator XR, https://operatorxr.com/, 2025. Company website. Accessed: 2025-12-19.

[19] HGXR, HOLOFORCE BLUE, https://hgxr.com/holoforce-blue/, 2025. Company webpage. Accessed: 2025-12-19.

[20] Verlag Dashöfer GmbH, VR EasySpeech – Erläuterung der Parameter, Technical Report, Verlag Dashöfer GmbH, Hamburg, Germany, 2024. URL: https://static.dashoefer.de/download/vr/whitepaper_auswertungsparameter_20240717.pdf.
[21] M. Murtinger, J. C. Uhl, L. M. Atzmüller, G. Regal, M. Roither, Sound of the Police—Virtual Reality Training for Police Communication for High-Stress Operations, Multimodal Technologies and Interaction 8 (2024) 46. doi:10.3390/mti8060046.
[22] J. E. Muñoz, J. A. Lavoie, A. T. Pope, Psychophysiological insights and user perspectives: enhancing police de-escalation skills through full-body vr training, Frontiers in Psychology 15 (2024) 1390677.
[23] T. Baur, I. Damian, P. Gebhard, K. Porayska-Pomsta, E. André, A job interview simulation: Social cue-based interaction with a virtual character, in: 2013 International Conference on Social Computing, 2013, pp. 220–227. doi:10.1109/SocialCom.2013.39.
[24] T. Baur, I. Damian, F. Lingenfelser, J. Wagner, E. André, NovA: Automated Analysis of Nonverbal Signals in Social Interactions, in: D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. A. Salah, H. Hung, O. Aran, H. Gunes (Eds.), Human Behavior Understanding, volume 8212, Springer International Publishing, Cham, 2013, pp. 160–171. doi:10.1007/978-3-319-02714-2_14, series Title: Lecture Notes in Computer Science.
[25] P. Bartyzel, M. Igras-Cybulska, D. Hekiert, M. Majdak, G. Łukawski, T. Bohné, S. Tadeja, Exploring user reception of speech-controlled virtual reality environment for voice and public speaking training, Computers & Graphics 126 (2025) 104160. doi:10.1016/j.cag.2024.104160.
[26] J. Tauscher, A. Witt, S. Bosse, et al., Exploring neural and peripheral physiological correlates of simulator sickness. comput anima virtual worlds 31: e1953, 2020.
[27] Y. Chen, T. Stephani, M. T. Bagdasarian, A. Hilsmann, P. Eisert, A. Villringer, S. Bosse, M. Gaebler, V. V. Nikulin, Realness of face images can be decoded from non-linear modulation of eeg responses, Scientific Reports 14 (2024) 5683.
[28] G. Chanel, C. Rebetez, M. Bétrancourt, T. Pun, Emotion assessment from physiological signals for adaptation of game difficulty, IEEE Trans. Syst. Man Cybern. A:Syst. Hum. 41 (2011) 1052–1063. doi:10.1109/TSMCA.2011.2116000.
[29] H. Gjoreski, I. I. Mavridou, M. Fatoorechi, I. Kiprijanovska, M. Gjoreski, G. Cox, C. Nduka, emteqpro: Face-mounted mask for emotion recognition and affective computing, in: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, UbiComp/ISWC '21 Adjunct, Association for Computing Machinery, New York, NY, USA, 2021, p. 23–25. doi:10.1145/3460418.3479276.
[30] M. Gnacek, J. Broulidakis, I. Mavridou, M. Fatoorechi, E. Seiss, T. Kostoulas, E. Balaguer-Ballester, I. Kiprijanovska, C. Rosten, C. Nduka, Emteqpro—fully integrated biometric sensing array for non-invasive biomedical research in virtual reality, Frontiers in Virtual Reality 3 (2022) 781218. doi:10.3389/frvir.2022.781218.
[31] I. Kiprijanovska, B. Sazdov, M. Majstoroski, S. Stankoski, M. Gjoreski, C. Nduka, H. Gjoreski, Facial expression recognition using facial mask with emg sensors., in: VR4Health@ MUM, 2022, pp. 23–28.
[32] M. Gjoreski, I. Kiprijanovska, S. Stankoski, I. Mavridou, M. J. Broulidakis, H. Gjoreski, C. Nduka, Facial emg sensing for monitoring affect using a wearable device, Scientific Reports 12 (2022) 16876. doi:10.1038/s41598-022-21456-1.
[33] S. Hickson, N. Dufour, A. Sud, V. Kwatra, I. Essa, Eyemotion: Classifying facial expressions in vr using eye-tracking cameras, in: Proc. IEEE WACV, 2019, pp. 1626–1635. doi:10.1109/WACV.2019.00178.
[34] M. Murakami, K. Kikui, K. Suzuki, F. Nakamura, M. Fukuoka, K. Masai, Y. Sugiura, M. Sugimoto, Affectivehmd: facial expression recognition in head mounted display using embedded photo reflective sensors, in: ACM SIGGRAPH 2019 Emerging Technologies, SIGGRAPH '19, Association for Computing Machinery, New York, NY, USA, 2019. doi:10.1145/3305367.3335039.
[35] N. Numan, F. t. Haar, P. Cesar, Generative rgb-d face completion for head-mounted display removal,

in: Proc. IEEE VRW, 2021, pp. 109–116. doi:10.1109/VRW52623.2021.00028.

[36] K. Tomotaki-Dawoud, B. Nierula, F. T. Siewe, T. Koch, D. J. Meyer, A. Bock, M. Heinze, D. Knuth, D. Martin, J. Schander, A. Hilsmann, P. Eisert, S. Bosse, Multi-view gesture recognition in conflict situations, in: 2024 International Symposium on Multimedia (ISM), 2024, pp. 267–268. doi:10.1109/ISM63611.2024.00060.

[37] P. L. Indrasiri, B. Kashyap, C. Kolambahewage, B. Nakisa, K. Ijaz, P. N. Pathirana, Vr based emotion recognition using deep multimodal fusion with biosignals across multiple anatomical domains (2024). URL: https://arxiv.org/abs/2412.02283. arXiv:2412.02283.

[38] E. M. Polo, F. Iacomi, A. V. Rey, D. Ferraris, A. Paglialonga, R. Barbieri, Advancing emotion recognition with virtual reality: A multimodal approach using physiological signals and machine learning, Computers in Biology and Medicine 193 (2025) 110310. doi:https://doi.org/10.1016/j.compbiomed.2025.110310.

[39] B. Nierula, K. Tomotaki-Dawoud, M. Akguel, M. T. Lafci, D. Przewozny, A. Hilsmann, P. Eisert, S. Bosse, Occlusion-robust multimodal emotion recognition in VR via fusion of facial images and EMG, 2026. Accepted at ACM IUI 2026 Workshop SHAPEXR.

[40] C. Kothe, S. Y. Shirazi, T. Stenner, D. Medine, C. Boulay, M. I. Grivich, F. Artoni, T. Mullen, A. Delorme, S. Makeig, The lab streaming layer for synchronized multimodal recording, Imaging Neuroscience 3 (2025) IMAG.a.136. URL: https://doi.org/10.1162/IMAG.a.136. doi:10.1162/IMAG.a.136, open Access.

[41] LSL Developers, Lab Streaming Layer (LSL), https://github.com/sccn/labstreaminglayer, 2025. GitHub repository. Accessed: 2025-12-19.

[42] SYSTRAN SA, Faster Whisper transcription with CTranslate2, https://github.com/SYSTRAN/faster-whisper, 2025. GitHub repository. Accessed: 2025-12-19.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, 2022. doi:10.48550/arXiv.2212.04356.

[44] D. Wiśniewski, Z. Rostek, A. Nowakowski, Fame-mt dataset: Formality awareness made easy for machine translation purposes., arXiv preprint arXiv:2405.11942. (2024).

[45] M. Nadejde, A. Currey, B. Hsu, X. Niu, M. Federico, G. Dinu, Cocoa-mt: A dataset and benchmark for contrastive controlled mt with application to formality., In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 616-632). (2022).

[46] H. Asghari, F. Hewett, Hiig at germeval 2022: Best of both worlds ensemble for automatic text complexity assessment, In Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text, pages 15–20 (2022).

[47] B. Naderi, S. Mohtaj, K. Ensikat, S. Möller, Subjective assessment of text complexity: A dataset for german language., arXiv preprint arXiv:1904.07733. (2019).

[48] D. Wu, T. D. Parsons, S. S. Narayanan, Acoustic feature analysis in speech emotion primitives estimation, in: Interspeech 2010, 2010, pp. 785–788. doi:10.21437/Interspeech.2010-285.

[49] ISO 532-1:2017, Acoustics - Methods for calculating loudness. Part 1, Zwicker method, Technical Report, International Organization for Standardization, 2017.

[50] G. F. Coop, MOSQITO, 2025. doi:10.5281/zenodo.10629475.

[51] M. Mauch, S. Dixon, Pyin: A fundamental frequency estimator using probabilistic threshold distributions, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 659–663. doi:10.1109/ICASSP.2014.6853678.

[52] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, B. W. Schuller, Dawn of the transformer era in speech emotion recognition: Closing the valence gap, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 10745–10759. doi:10.1109/tpami.2023.3263585.

[53] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. L. Zhu, F. Zhang, M. Grundmann, Blazepose: On-device real-time body pose tracking, ArXiv abs/2006.10204 (2020).

[54] I. Grishchenko, V. Bazarevsky, A. Zanfir, E. G. Bazavan, M. Zanfir, R. Yee, K. Raveendran, M. Zhdanovich, M. Grundmann, C. Sminchisescu, Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation, 2022. arXiv:2206.11678.

[55] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, C. Sminchisescu, Ghum & ghuml: Generative 3d human shape and articulated pose models, in: IEEE-CVPR, 2020, pp. 6183–6192. doi:10.1109/CVPR42600.2020.00622.

[56] P. Ekman, An argument for basic emotions, Cognition and Emotion 6 (1992) 169–200. Publisher: Routledge.

[57] V. V. Nikulin, G. Nolte, G. Curio, A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition, NeuroImage 55 (2011) 1528–1535. doi:10.1016/j.neuroimage.2011.01.057.

[58] S. Dähne, F. C. Meinecke, S. Haufe, J. Höhne, M. Tangermann, K.-R. Müller, V. V. Nikulin, SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters, NeuroImage 86 (2014) 111–122. doi:10.1016/j.neuroimage.2013.07.079.

[59] S. M. Hofmann, F. Klotzsche, A. Mariola, V. Nikulin, A. Villringer, M. Gaebler, Decoding subjective emotional arousal from EEG during an immersive virtual reality experience, eLife 10 (2021) e64812. doi:10.7554/eLife.64812.

[60] B. Nierula, M. T. Lafci, A. Melnik, M. Akgül, F. T. Siewe, S. Bosse, Differential Physiological Responses to Proxemic and Facial Threats in Virtual Avatar Interactions, 2025. doi:10.48550/ARXIV.2508.10586.

[61] R. Flewitt, S. Price, T. Korkiakangas, Multimodality: Methodological explorations, Qualitative Research 19 (2018) 3–6. doi:10.1177/1468794118817414.

[62] H. Blumer, Symbolic interactionism: Perspective and method, Prentice Hall, Englewood Cliffs, NJ, 1969.

[63] L. D. Keesman, D. Weenink, Feel it coming: Situational turning points in police-civilian encounters, Historical Social Research 47 (2022) 88–110. doi:10.12759/hsr.47.2022.04.

[64] H. M. Sunde, How does it end well? an interview study of police officers' perceptions of de-escalation, Nordic Journal of Studies in Policing 11 (2024) 1–21. URL: https://doi.org/10.18261/njsp.11.1.1. doi:10.18261/njsp.11.1.1.

[65] J. Hu, L. Mathur, P. P. Liang, L.-P. Morency, Openface 3.0: A lightweight multitask system for comprehensive facial behavior analysis, arXiv preprint arXiv:2506.02891 (2025).

[66] B. Nierula, M. T. Lafci, A. Melnik, E. Gaudinot, N. Karuzin, S. Bosse, Personal space in virtual reality interactions assessed with electroencephalography and skin conductance, in: Neuroscience 2025, San Diego, California, United States, 2025.

[67] M. T. Lafci, B. Nierula, D. Damar, S. Bosse, Too Close for Comfort? Investigating Virtual Professor Distance and Student Learning in VR, in: IEEE SMC, Vienna, 2025.