The Color of Emotions: a Culturally-Aware Visual Emotional Regulation Architecture for Social Robotics

Selena Magno[1,†], Arianna Pipitone[1,2,*,†]

[1]AIRHLab - Artificial Intelligence and Robotics for Humanities Lab, Department of Humanities, Building 12, University of Palermo
[2]ICAR - CNR National Research Council, Palermo site

Abstract

Socially assistive robots increasingly rely on Facial Expression Recognition (FER) to perceive the affective state of their users, yet most systems stop at recognition and react in a culturally "one-size-fits-all" manner. Because they are trained predominantly on WEIRD (Western, Educated, Industrialized, Rich, Democratic) data, they treat the symbolic meaning of colors and emotions as universal, even though this meaning is a culturally learned construct. We present the *Visual Emotional Regulation System* (VERS), a modular, device-agnostic architecture that dissociates *perception* (treated, as a deliberate operational assumption, as largely universal) from *signification* (which is culture-specific). A perceptual module recognizes the user's emotion from facial expressions; a cultural-clustering module maps a declared nationality to a macro-cluster; and a chromatic-actuation module composes a regulating color by passing a culture-free regulatory intent through the user's culture-specific symbolic code, verbalized through an *inner-speech* layer that grants cognitive transparency, rendering each cultural assumption inspectable and revisable.

A systematic, data-centric comparison of five vision backbones across three datasets shows that data quality dominates architectural depth, with a Swin Transformer trained on a large, class-balanced corpus clearly outperforming deeper but data-starved alternatives. Crucially, the regulating color is not read from a lookup table but composed at each cycle from a culture-free intent and a culture-specific symbolic code, so that identical detected emotions yield culturally-differentiated colors, each delivered together with its verbalized rationale. A real-time prototype illustrates this behavior in a preliminary, two-participant proof-of-concept A/B demonstration, in which the same detected emotion is regulated with different, culturally-licensed hues; a protocol for large-scale cross-cultural validation is defined and left to future work. We chart the limits of macro-cluster reasoning and outline a path toward micro-adaptive, biometrically closed-loop regulation. VERS thus reframes affective robotics as a problem of *meaning*, not merely of *recognition*.

Keywords

Culture-Aware Emotion Regulation, Chromatic Affective Regulation, Facial Emotion Recognition, Robot's inner speech, Robot's cognitive transparency, Affective HRI, Social Robotics

CEUR-WS.org/Vol-4314/ReDH8.pdf

1. Introduction

The evolution of Artificial Intelligence (AI) and Computer Vision is rapidly reshaping Human-Robot Interaction (HRI). Where the historical goal was simply to execute instructions, the field now aims at machines endowed with a form of emotional intelligence: machines that, beyond seeing objects or faces, understand and react to the nuances of human behavior [1]. As robots move into assistive, therapeutic and educational environments - the domain of *Socially Assistive Robotics* (SAR) [2] - mere facial recognition is no longer sufficient; artificial emotional intelligence becomes a functional requirement.

The current state of the art offers highly performant Facial Expression Recognition (FER) systems, but it exhibits a structural gap: these systems *recognize* emotions without *responding* in a culturally adequate way. Most models inherit the so-called WEIRD bias (Western, Educated, Industrialized, Rich, Democratic) [3]: they are trained on datasets dominated by Western populations and implicitly assume that facial expressions, the perception of emotions and the semantics of color are universal, when in

Rethinking the future of Digital Humanities 2026 workshop, in conjunction with AIxIA 2026
*Corresponding author.
†These authors contributed equally.
arianna.pipitone@unipa.it (A. Pipitone)
0000-0003-2388-5887 (A. Pipitone)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

fact the latter two are culturally learned constructs. A robot that ignores the cultural context of its user therefore risks delivering ineffective, or even counterproductive, feedback.

Our proposal rests on a foundational theoretical distinction between *perception* and *signification*. Since the work of Berlin and Kay on basic color terms [4] and Ekman on basic emotions [5], there is broad evidence of a universal regularity in the perception of primary stimuli: people from different cultures discriminate chromatic *foci* and basic facial expressions in substantially the same way (a universalist reading that, for facial expressions, is contested [6] and that we adopt as a deliberate operational assumption, see Section 3). Cultural variability, however, emerges forcefully in the association of *symbolic* meanings, and this divergence increases with the complexity of the stimulus: as one moves away from primary colors or instinctive emotions, the social constraint becomes predominant. The system proposed here sits exactly at this interface. It does not attempt to correct perception but to manage symbolic variability, adapting the machine's response to the user's cultural code, in line with Color-in-Context theory [7] and with the notion of Cultural Awareness in HRI [8, 9].

In this paper we introduce the *Visual Emotional Regulation System* (VERS), which locates the user within a cultural macro-cluster from a declared nationality and answers the detected emotion with a regulating color (soothing negative states, reinforcing positive ones) adapted to its meaning in the user's culture, while an inner-speech layer verbalizes the decision. The contributions are the following:

- a modular, device-agnostic architecture that explicitly dissociates universal perception from culture-specific actuation, and couples it to a transparent, verbalized decision process;
- a systematic, data-centric experimental comparison of CNN, pure-Transformer and hybrid backbones across three FER datasets, showing that data quality dominates architectural depth and identifying a reliable perceptual engine for HRI;
- a real-time prototype, a protocol for cross-cultural validation, and a preliminary proof-of-concept A/B demonstration illustrating that identical detected emotions can yield culturally-differentiated chromatic responses;
- a critical analysis of the limits of macro-cluster reasoning, distinguishing invariant from high-variance color-emotion associations, and a research agenda toward micro-adaptive, closed-loop regulation.

The remainder of the paper is organized as follows. Section 2 reviews the background; Section 3 presents the architecture; Section 4 describes the prototype and the selection of its perceptual model; Section 5 presents the validation protocol and a proof of concept; Sections 6 and 7 discuss limits and conclude.

2. Background and Related Work

2.1. Color Psychology, Context, and Cross-Cultural Semantics

Color is far more than a visual stimulus: it conveys symbolic meanings that influence emotion, cognition, and behavior. According to Color-in-Context theory [7], the psychological impact of a color emerges from the interaction between biologically grounded perception and context-dependent learned associations. While humans share similar perceptual mechanisms for discriminating colors, their symbolic interpretation is strongly shaped by cultural norms and social experience, a tension already highlighted by early cross-cultural work on the affective connotations of color [10] and on the universality and evolution of basic color terms [4].

Cross-cultural research confirms that color semantics are neither completely universal nor entirely culture-dependent. A recent large-scale survey involving thousands of participants across dozens of nations and native languages found robust universal color–emotion associations that are nonetheless further modulated by linguistic and geographic proximity [11]. Some color–emotion associations, such as red with love or anger and black with sadness, appear remarkably stable across geographically and linguistically distant populations, suggesting a partially biological origin. Conversely, colors

such as white, yellow, and purple exhibit substantial semantic variability, assuming different or even opposite meanings depending on the cultural context (e.g., purity versus mourning, prosperity versus jealousy), a variability also documented in cross-national studies of color meanings and preferences [12]. Consequently, identical chromatic stimuli may produce different emotional interpretations across users.

This distinction has direct implications for Human–Robot Interaction. When colors are employed as part of an affective regulation strategy, their effectiveness depends not only on perceptual salience but also on their cultural interpretation. The principles of Cultural Awareness in robotics [8] and the broader notion of cultural robustness [9] therefore argue against designing robots for a hypothetical "universal user". Instead, socially intelligent systems should adapt their communicative strategies to the cultural background of the individual with whom they interact.

These observations motivate the central premise of this work: while emotion perception may rely on largely universal mechanisms, the selection of an appropriate chromatic response should account for the user's cultural framework.

2.2. Social Robotics and Affective Computing

Socially Assistive Robotics (SAR) extends traditional robotics by placing social interaction at the center of the robot's functionality. Unlike conventional robotic systems, whose primary objective is task execution, socially assistive robots operate in therapeutic, educational, and healthcare environments where successful interaction depends on the robot's ability to perceive, interpret, and respond appropriately to human affective states [1, 2]. This capability is grounded in the field of affective computing, originally articulated as the design of systems able to perceive, interpret, and simulate human affect [13], and more recently systematized in comprehensive reviews of emotion models, databases, and computational advances [14].

Recent advances in affective computing have enabled robots to move beyond simple emotion recognition toward adaptive behavior generation. Modern cognitive architectures increasingly combine affective information with contextual knowledge to personalize robot actions in real time, often relying on AI planning techniques to dynamically select interaction strategies [15]. However, personalization is typically driven by the detected emotional state alone, whereas the cultural interpretation of the robot's expressive signals remains largely unexplored.

For systems employing visual emotional cues such as color, this limitation becomes particularly relevant. Color is an inexpensive, flexible, and perceptually salient channel for conveying affect, and prior work in HRI has shown that colored light and chromatic facial cues can meaningfully shape how users attribute emotions to a robot [16, 17, 18]. Nonetheless, a regulating strategy that is effective for one cultural group may be ambiguous or even counterproductive for another. Therefore, affective adaptation should consider not only how emotions are perceived, but also how the robot's responses are interpreted by users from different cultural backgrounds.

2.3. Inner Speech and Cognitive Transparency

Adaptive behavior alone is insufficient to establish effective human-robot collaboration. As robots become increasingly autonomous, users must also understand the reasoning that guides their decisions. This requirement has motivated the growing adoption of Explainable AI (XAI) and Interpretable AI (IAI), which seek to make robot behavior more transparent and understandable, mitigating the opacity often associated with deep learning systems [19].

Within social robotics, one promising approach consists of equipping the robot with an explicit *inner speech*, namely an externalized internal dialogue through which the robot verbalizes its reasoning process [20, 21]. Rather than serving merely as a debugging tool, inner speech enables users to follow the cognitive steps leading to a decision, reducing ambiguity and fostering trust during interaction. More recently, this paradigm has been extended to affective reasoning, allowing robots to explicitly communicate not only what they intend to do, but also why a particular emotional response has been selected [22].

This capability is particularly relevant in culturally adaptive systems. When different users receive different regulating colors despite displaying the same detected emotion, transparency becomes essential to avoid the impression of arbitrary behavior and to make the adaptation process understandable.

2.4. Facial Expression Recognition for Affective Perception

A fundamental issue for culturally adaptive responses is the estimation of the user's current affective state. Facial Expression Recognition (FER) has therefore become one of the most widely adopted perception mechanisms in affective HRI, enabling robots to infer emotions from facial cues in a non-invasive and real-time manner. The field has been extensively systematized in recent surveys of deep-learning FER techniques, which map its architectures, benchmark datasets, and open challenges [23, 24].

Convolutional Neural Networks (CNNs) have traditionally dominated FER thanks to their strong inductive biases, including locality and translation invariance, which allow efficient learning from relatively limited datasets. More recently, Vision Transformers (ViTs) [25] have demonstrated superior capability in capturing long-range dependencies through self-attention, producing richer global representations at the cost of higher computational and data requirements. The Swin Transformer [26] addresses this trade-off by combining hierarchical feature extraction with window-based self-attention, achieving an effective balance between recognition performance and computational efficiency. Transformer-based architectures have also been adapted specifically to FER, learning expression-related representations while mitigating the data demands typical of attention-based models [27]. These characteristics make such models particularly suitable for real-time HRI, where robust perception must coexist with limited computational resources.

Within VERS, FER constitutes the perception layer of the architecture, providing the emotional state that drives the subsequent culture-aware regulation process. Since the effectiveness of the overall system depends on the robustness of this perceptual module, different deep-learning backbones are experimentally evaluated and compared in the implementation phase to identify the most suitable model for real-time deployment.

3. The Visual Emotional Regulation System

3.1. Design Principles

VERS is a modular, *device-agnostic* software architecture that runs on any device with standard compute, a camera and a screen. Two commitments shape the system. First, it is a *closed-loop* architecture: whereas conventional facial-recognition pipelines terminate at classification, VERS adds an active response stage that continuously modulates the user's affective state and re-reads the face, treating regulation as an ongoing control problem rather than a one-shot inference. Second, and more distinctively, it dissociates *universal perception* from *culture-specific actuation*: where standard affective systems map an emotion onto a fixed reaction, VERS interposes a *cultural filter* between sensing and acting, so that the same detected emotion yields different regulating colors according to the user's background.

Functionally, VERS unfolds as a *sensing-thinking-acting* loop of four cooperating modules (Figure 1), preceded by an offline training phase in which the perceptual backbone learns the weights required for emotion recognition; at inference time the biometric datum (emotion) and the contextual datum (culture) converge to produce the regulating output. The loop is bound together by an *inner-speech* layer [21, 20] that recalls the cultural knowledge from declarative memory and rehearses it into a first-person justification, and that also utters *motivational* (negative emotions) or *reinforcing* (positive emotions) messages, so that regulation is delivered both visually and verbally. Since actuation reduces to a screen and a voice, it transfers to any screen-equipped embodiment, such as Pepper's chest display. The reference deployment is *client-server*: a workstation carries vision and reasoning, while the embodied agent acts as social interface for acquisition and actuation.

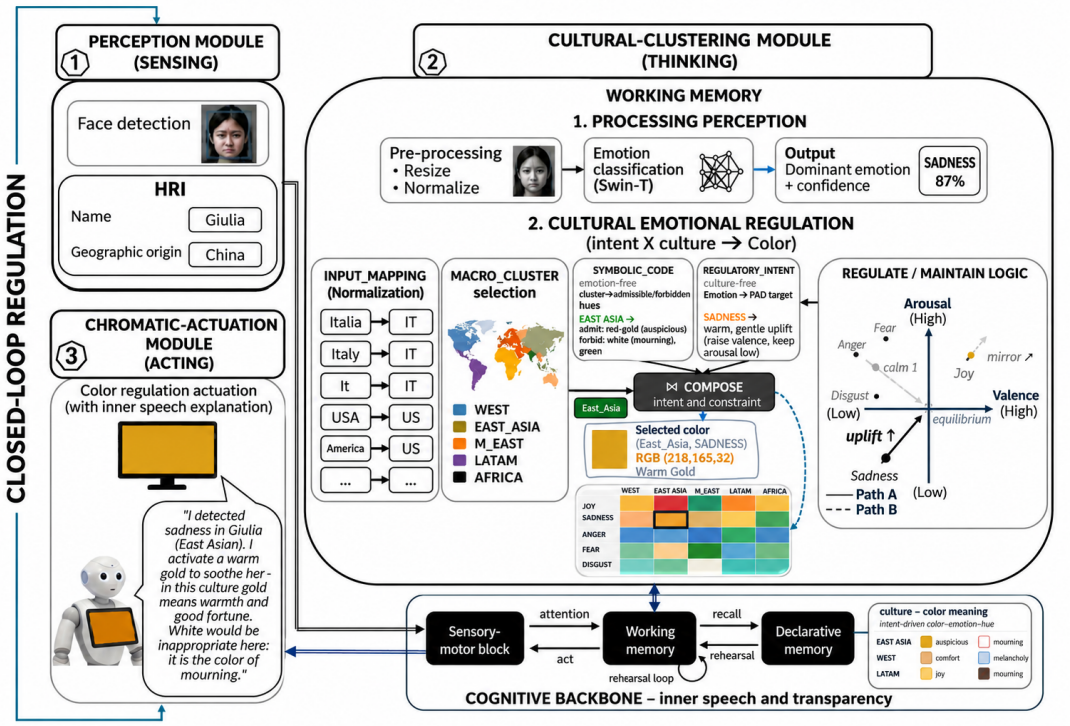


Figure 1: High-level VERS architecture: perception (sensing) is dissociated from culture-specific actuation (acting); the thinking stage performs cultural clustering and color selection; the inner-speech layer verbalizes every stage.

3.2. Perception Module (Sensing)

The perception module supplies the loop with an objective, culturally neutral acquisition of the user's state. The webcam stream is captured frame by frame and a face detector isolates the Region of Interest (ROI); alongside this biometric channel, the module also handles the interactive exchange through which the user declares a name and a geographic origin, so that both the facial datum and the contextual datum enter the loop at the sensing stage.

This output enters working memory and is rehearsed as a first-person observation (e.g. *"I am looking at Giulia's face; I'm trying to process the muscular configuration for estimating emotion"*), mirrored on screen by a bounding box and the predicted label.

3.3. Cultural-Clustering Module (Thinking)

This module fuses biometric and cultural information within working memory: it runs the emotion classification on the ROI, yielding the dominant emotion and its confidence, and retrieves the declared origin, stored in declarative memory. Anchoring detection in Ekman's basic emotions [5] is a deliberate choice rather than an unexamined premise. Their universality is contested: constructionist accounts argue that emotion categories lack fixed facial signatures [28], and cross-cultural studies report culture-specific expressive patterns and recognition rates below universalist predictions [29, 30, 31, 6]. We adopt the basic-emotion taxonomy pragmatically, as the label space of the available large-scale FER corpora; the perception/signification boundary is thus an engineering idealization, and the cultural variability located in perception itself is not modeled by the sensing stage. Assuming that sensing has identified the emotion and that focal colors are perceived coherently [4], the module's task is to select a color that *regulates* the detected emotion while respecting the local symbolic code.

Crucially, this regulating color is not read off a cell but *composed* from two independently grounded axes. The first is emotion-driven and *culture-free*: the detected emotion fixes a *regulatory intent*, a target region in valence-arousal Russell's space (*circumplex model*) [32] specifying *what kind* of response is

Table 1
CHROMA_REGULATION as the cross-tabulation of the intent × constraint composition: each cell is the RGB stimulus the COMPOSE step selects to steer the detected emotion toward equilibrium: the regulatory intent picks the target, the cluster's symbolic code filters the admissible hues. Entries were authored manually from the cited literature (see text) and have not yet undergone inter-rater validation.

Emotion	Regulation intent	WEST	EAST_ASIA	M_EAST	LATAM	AFRICA[†]
Joy	Reinforce (mirror)	245,190,60	200,35,45	20,140,80	250,140,40	240,190,60
Sadness	Uplift (warm, gentle)	245,180,120	218,165,32	222,184,110	250,205,95	70,160,95
Anger	Calm (cool, low arousal)	70,140,205	95,160,205	55,125,190	55,170,110	70,140,190
Fear	Reassure (secure)	131,193,157	245,205,140	34,139,34	95,165,210	125,185,125
Disgust	Refresh (neutralize)	120,200,190	115,190,160	238,236,224	70,190,180	90,185,170

[†] The AFRICA macro-cluster aggregates highly heterogeneous color-symbol systems; its entries are the least reliable of the matrix and limit its generalizability.

required (for sadness, a warm, gently uplifting tone that raises valence while keeping arousal low). The second is culture-driven and *emotion-free*: the declared origin, resolved to a cultural cluster, recalls that cluster's *symbolic code*, i.e. the hue families its convention licenses or forbids - for EAST_ASIA, admit the auspicious red-gold family, forbid white (mourning) and green (the "green hat" idiom). At each cycle the module computes the *composition* of the two axes (the warmest, most uplifting hue the cluster does not proscribe) and the rehearsal process makes the two-part reasoning explicit, so that the thinking phase too produces inner speech (e.g. *"Giulia comes from China, hence East_Asia; she is sad, so I need a warm, uplifting tone; here gold is auspicious, whereas white would mean mourning"*).

Table 1 is the *cross-tabulation* of this operation (intent × constraint across five emotions and five clusters), not the rule itself. Entries are full RGB triples, so two cells of the same hue family can still differ. Keeping the axes separate yields *counterfactual portability* (swapping only the origin answers "and if Giulia were Western?") and *graceful degradation*: for an under-specified cluster or an unmapped country, the system can regulate on the culture-free intent alone and *say so*.

The matrix was hand encoded. The admissible and forbidden hue *families* and the regulatory intents are drawn from the cited literature [7, 11, 12, 32]. The specific RGB triple within a family is derived by applying the hue–saturation–brightness correlates of affect reported by Valdez and Mehrabian [33] (brightness raising pleasure, saturation raising arousal) to the target valence–arousal region; the resulting triples are therefore rule-based rather than individually attested by the cultural sources.

A central design decision is that the module reasons over cultural *macro-regions* rather than individual countries, grounded in research showing that nations recurrently group into a small number of stable clusters defined by shared values, language, religion, and geography [34, 35]. Five data structures in declarative memory implement and factor the decision:

- INPUT_MAPPING normalizes heterogeneous input ("Italia", "Italy", "IT") into ISO country codes;
- COUNTRY_TO_CLUSTER performs the cultural lookup onto the five macro-clusters (WEST, EAST_ASIA, M_EAST, LATAM, AFRICA);
- SYMBOLIC_CODE maps each cluster to its *emotion-free* chromatic convention, i.e. the hue families it admits or forbids;
- REGULATORY_INTENT maps the emotion to a *culture-free* valence–arousal target through the *regulate/maintain* logic: the sign of the valence selects the strategy. Along *Path A* (negative valence) the system opposes the state: anger (high arousal) is met with cool, low-arousal tones, sadness (low arousal) with warm, luminous but non-aggressive ones. Along *Path B* (positive valence, e.g. joy) it mirrors the state to reinforce well-being;
- CHROMA_REGULATION is the *derived* cross-tabulation: for each cluster × emotion, the RGB triple returned when the intent is satisfied under the constraint (e.g. sadness in EAST_ASIA yields a warm gold, not white or green).

Thus, for a *sad* user the intent is invariant, but the admissible hue varies by cluster: a warm peach for WEST (avoiding melancholic blue), a warm gold for EAST_ASIA (excluding white and green), a warm yellow for LATAM (excluding mourning purple and black).

3.4. Chromatic-Actuation Module (Acting)

The regulating color floods the full display, closing the feedback loop: as the detected expression changes, the color adapts. Colored light and full-field chromatic displays have been shown to shape affective response and emotion attribution in HRI [16, 17]. As the color is displayed, the rehearsal process emits an utterance naming the act and its purpose (e.g. *"I now bathe the scene in warm gold to soothe you"*), so that the user both *sees* the color and *hears* why it was chosen.

3.5. Inner Speech as the Cognitive Backbone

The inner-speech layer is not an explanatory add-on but the place where the cultural knowledge resides and decisions are formed. The SYMBOLIC_CODE, the declared origin and CHROMA_REGULATION form the agent's *declarative memory*; the detected emotion, mapped through REGULATORY_INTENT, is composed there with the cultural constraint; and the *rehearsal* process fixes the regulating color while verbalizing its two-part reason. Since selection and explanation come from a single recall-compose-rehearse cycle, each module leaves a first-person trace of *what* was perceived, *how* it was interpreted and *why* a color was chosen: VERS is transparent by construction rather than through a post-hoc report.

This transparency reaches beyond legibility. A bare correspondence table encodes culture as a closed, essentializing ontology, the very move for which national value-cluster cartographies are justly criticized, since it treats a nation as a homogeneous bundle of values. By factoring the decision into a culture-free intent and an explicit, separable symbolic code, and by surfacing that code as an *interpretive act* rather than a verdict (not *gold* but *"here gold is auspicious, whereas white would mean mourning"*) the rehearsal layer turns each cultural commitment into a claim that a user or critic can inspect, contest, and refuse. Transparency in VERS is therefore not cosmetic readability but the condition under which the system's cultural assumptions remain *revisable* rather than naturalized, and the macro-cluster approximations (coarse by construction, as the AFRICA entry shows) are treated as defeasible defaults rather than facts. The trace is always shown as an on-screen speech bubble and, when audio is available, also uttered aloud.

4. Prototype Implementation

The prototype instantiates the *sensing-thinking-acting* loop as a four-stage interactive session: (i) real-time emotional sensing through the perceptual backbone, which draws bounding boxes and labels on the live video; (ii) a short social exchange that collects the user's name and declared geographic origin; (iii) the culture-aware computation of the regulating color; and (iv) full-screen chromatic actuation with the inner-speech rationale. Speech synthesis runs asynchronously alongside video processing, so that the agent can "see" and "speak" at the same time. A stylized *Virtual Pepper* avatar provides the social interface (Figure 2); the physical robot, when present, behaves in the same way.

4.1. Perception Module: An Experimental Study

Every downstream decision is conditioned on the detected emotion, so a misread expression propagates unchanged to the actuator. Two constraints made a dedicated study necessary. First, the module must sustain a fluid real-time loop on accessible hardware, so accuracy has to be weighed against latency and memory footprint. Second, the *cost of error is asymmetric*: confusing emotions that call for opposite regulation (sadness, warm tones; anger, cool ones) is far more damaging than a generic drop in accuracy. The model was thus chosen against a safety threshold on the confusions that matter, not against an aggregate score.

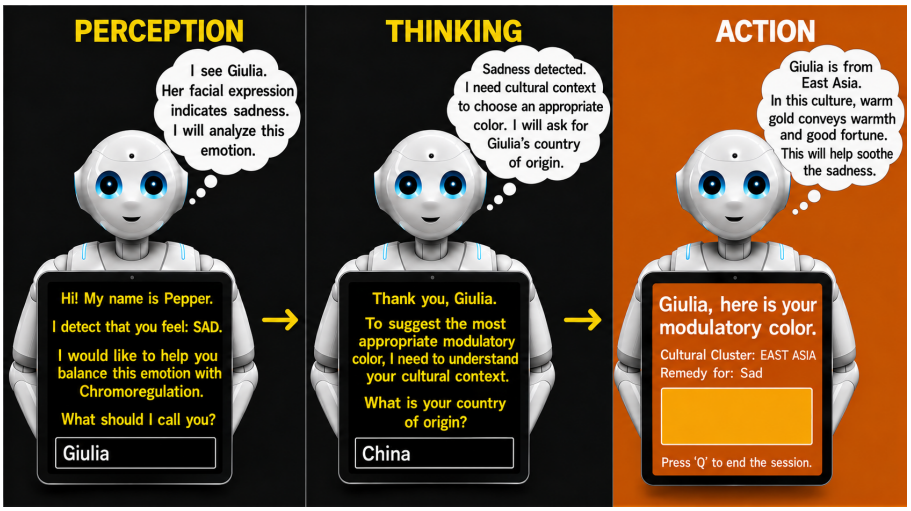


Figure 2: The Virtual Pepper social interface running on a screen through the three phases of the loop (sensing, thinking and acting), from real-time emotion recognition to the culture-aware chromatic actuation accompanied by its inner-speech rationale.

4.1.1. Experimental Setup

We compared five backbones contrasting the local inductive bias of CNNs with the global modeling of Transformers: *ResNet-18* ($\approx$ 11M parameters, lighter and less prone to overfitting on small or noisy data) and *ResNet-50* ($\approx$ 25M parameters, the industrial CNN benchmark); *ViT-Base* (pure global attention) together with its distilled variant *ViT-Tiny* (for resource-constrained scenarios); and the hybrid, hierarchical *Swin-T*.

These backbones were trained on three datasets of increasing quality and balance. *FER2013* [36] is the historical benchmark (35,887 grayscale 48×48 images over seven basic emotions), but it suffers from label noise (estimated human accuracy $\approx 65\%$) and strong class imbalance (e.g. 547 disgust vs. 8,989 happiness samples), which penalizes deep networks. *FERPlus* [37] keeps the same images but replaces single hard labels with crowd-sourced soft labels from ten annotators and adds finer classes such as *contempt* [38]. Finally, *FER2025* [39] is a large-scale, class-balanced corpus of over 1.5 million 75×75 images designed to remove the imbalance of FER2013 and to fully exploit modern Transformers, publicly released under a CC BY-NC 4.0 license[1]. FER2025 is a self-published resource by a single author: it has not undergone peer review, has seen limited adoption to date, and its label quality and demographic composition have not been independently validated. Since our best result is obtained on it, the figures below should be read with this caveat and confirmed on independently validated corpora.

Each dataset was split 70/20/10 (train/validation/test) with class stratification; mini-batch gradient descent used a batch size of 64. CNNs were trained on Google Colab, while the data-hungry Transformers exceeded its limits and were moved to Kaggle kernels (P100/T4 GPUs). To fit the available VRAM, a class-balanced subset of FER2025 was used for the final Swin-T training.

4.1.2. Results

Contrary to the expectation that deeper networks perform better, ResNet-50 on the noisy FER2013 struggled to exceed 60-65%: with its 25M parameters it tended to *fit the noise* rather than generalize. The lighter ResNet-18 on the clean FERPlus instead reached 81.14%, a reversal showing empirically that a shallower model on quality data beats a deeper model on poor data. Residual difficulties persisted only on minority classes (F1 of 0.33 for disgust, 0.29 for contempt), a direct consequence of class imbalance later resolved by FER2025.

Vision Transformers ran into a computational limit. A pure ViT on FER2013 reached only 72.44%:

[1] https://huggingface.co/datasets/imadhavan/FER2025, DOI: 10.57967/hf/8296.

Table 2
Summary of the experimental study. Data quality dominates architectural depth; the hybrid backbone on a balanced corpus is the only configuration that surpasses the 80% threshold.

Architecture	Dataset	Accuracy
ResNet-50 (CNN)	FER2013	60–65%
ResNet-18 (CNN)	FERPlus	81.14%
ViT (pure Transformer)	FER2013	72.44%
ViT (pure Transformer)	FER2025	failed (OOM)
ViT-Tiny	FER2025	low / underperforming
Swin-T (hybrid)	FER2013	70–72%
Swin-T (hybrid)	**FER2025**	$\mathbf{87.87}\%$

higher than ResNet-50 on the same data, yet below ResNet-18 on FERPlus, and marked by a *diverging gap*, training accuracy shooting toward 95% while validation stalled near 72% and validation loss began to rise, the classic overfitting pattern. The consequences were critical for HRI: the fear class scored only 0.57 F1, with many fear samples misclassified as sadness or anger. Attempts to train the ViT on FER2025 failed with Out-Of-Memory errors and session timeouts even after upgrading to Colab Pro, and the distilled ViT-Tiny trained but lost the capacity to capture the micro-expressions needed to separate similar emotions. Model compression alone was therefore not the answer.

The hybrid backbone provided the optimal balance. Even Swin-T on FER2013 remained at 70–72%, confirming that not even a state-of-the-art architecture can overcome low-quality input. Only the combination of the hybrid backbone with the balanced, high-resolution corpus (*Swin-T on FER2025*) broke the 80% barrier, reaching 87.87% test accuracy with stable convergence. Per-class F1-scores rose above 0.90 on historically hard classes (0.96 for disgust, 0.93 for fear), with only a mild residual confusion between static sadness and neutrality. Repeated runs with random re-shuffling of the splits stayed consistently above 85%, indicating genuine generalization rather than a fortunate split.

Table 2 summarizes the campaign. Beyond raw numbers, a qualitative real-time comparison exposed operational issues invisible to metrics alone: ResNet-50 showed unstable, low-confidence bounding boxes on noisy data, and the ViT introduced latency that made the interaction jerky, whereas Swin-T combined accuracy with fluid, responsive inference. This is what the safety threshold demands: Swin-T's local-plus-global self-attention proved superior at disambiguating subtle expressions (e.g. disgust vs. anger) where CNNs failed, and modern GPU-accelerated hardware processed frames in real time without interaction-breaking latency. Swin-T on FER2025 was therefore selected as the perceptual engine, the only configuration meeting the safety threshold required for reliable empathic interaction.

5. Cross-Cultural Validation Protocol and Proof of Concept

To assess VERS, we defined an evaluation protocol intended for large-scale administration; its application to a broad, culturally balanced cohort is future work. In this paper the protocol is applied only in a two-participant proof of concept, which illustrates the behavior of the cultural module and the feasibility of the procedure, but does not constitute a validation of it.

5.1. Evaluation Protocol

The planned study will involve approximately $N = 50$ participants who interact with the platform. Two five-point ratings together with open verbal feedback, correlating the participant's declared cultural background with their assessment of the system's response, will enable VERS assessment; Table 3 summarizes the fields and scales the protocol will record. *Cultural congruence* ranges from "totally inappropriate/offensive" (1) to "perfectly appropriate" (5) and captures whether the chosen colour respects the user's symbolic code; *emotional efficacy* ranges from "no effect/worsening" (1) to "sense of well-being/calm" (5) and captures the regulation's felt effect. The open feedback surfaces discrepancies

Table 3
Data-collection protocol for subjective validation: the two ordinal scales and the open feedback recorded for each interaction.

Field	Scale	Captures
Cultural congruence	1–5	Appropriateness to the symbolic code
Emotional efficacy	1–5	Felt regulatory effect
Verbal feedback	Open	Rationale and model discrepancies

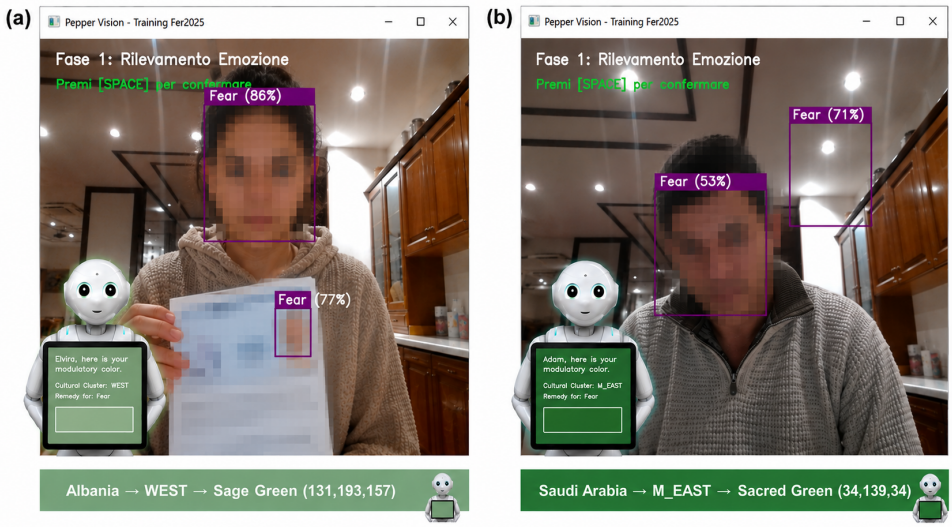


Figure 3: Proof-of-concept A/B demonstration for an identical detected emotion (*fear*): (a) an Albanian user (West) receives a soft Sage Green, RGB $(131, 193, 157)$; (b) a Saudi user (M_East) receives a Sacred Green, RGB $(34, 139, 34)$. *Anecdotal ratings* ($n = 2$, one color per participant; not statistically interpretable): Albanian user, cultural congruence 5 and emotional efficacy 4; Saudi user, cultural congruence 4 and emotional efficacy 4.

between the theoretical model and lived experience that the numerical scales alone would miss. The protocol is deliberately model-agnostic and scale-independent (the same grid applies unchanged to any cluster, emotion and cohort size) so that it can be administered as-is to a large, culturally balanced sample.

5.2. Proof of Concept: an A/B Demonstration

As a proof of concept, we applied the protocol to a controlled A/B demonstration with two participants, one Albanian and one Saudi, each exposed to a single color. The demonstration holds the detected emotion fixed (*fear*) and varies only the declared nationality. The design is deliberate: if the system implemented a rigid, emotion-only mapping, both users would receive an identical colour, so any difference in the output can be attributed solely to the cultural context. As reported in Figure 3, the algorithm discriminated the two contexts and returned two renderings from the same green family.

Both users receive a green, consistent with the regulation logic of Section 3: fear is a high-arousal, negative-valence state, so Path A calls for a calming, non-aggressive tone that restores a sense of safety, and green satisfies this requirement while being culturally positive in both clusters. What differs is the rendering, tuned to each local code: for the M_East cluster a deep, saturated "Sacred Green" evokes protection and spirituality, countering fear with solidity; for the West cluster a soft, desaturated "Sage Green" evokes nature and grounding, reassuring without visual aggression.

The distinction is not cosmetic. The two users receive different RGB triples and two different triples are two different colours, however similar their names: the stimulus produced by flooding a display with the former is not the one produced by the latter. That a shared hue name ("green") resolves to distinct, culturally-selected values is precisely the behaviour a rigid emotion-only mapping could never

produce, since such a mapping would return one fixed colour for fear regardless of who is watching.

This case is intentionally conservative: it isolates cultural tuning *within* a shared hue family, the subtlest form the adaptation can take. The same mechanism produces full *cross-hue* divergence when the local codes disagree more sharply, for instance sadness regulated with a warm gold in the East_Asia cluster versus a warm peach in the West, or the outright exclusion of taboo hues such as white, as tabulated in the regulation matrix (Table 1). Only the within-family case was presented to participants in this proof of concept; the cross-hue cases are produced by the implemented pipeline but have not yet been assessed with users. Across both levels of variation the intended behavior is the same: the system does not merely apply universal color psychophysiology but overlays a semantic filter aimed at a safe, empathic interaction.

Both participants responded favorably (anecdotal ratings are reported in the caption of Figure 3); their verbal feedback linked the tuned tones to felt "protection" and "relaxation" and described the robot's spoken rationale as "scientific and reassuring". With two participants, each exposed to a single color, these responses show that the pipeline behaves as designed and that the protocol can be administered end to end; they do not establish the cultural congruence or the regulatory efficacy of the module, which require the planned cohort study against a culturally neutral baseline.

6. Discussion, Limitations and Future Work

The results are consistent with the initial hypotheses, but with different degrees of evidential support. The most solid is empirical: Swin-T on FER2025 is a reliable perceptual engine (87.87%), although this figure is obtained on a corpus that has not yet been peer reviewed. The prototype closes the recognition→decision→color loop in real time on standard hardware. The proof of concept illustrates that cultural clustering can differentiate the chromatic response to the same emotion and exclude "taboo" colors; this is a demonstration of designed behavior, not an empirical validation. The perception/signification distinction therefore remains, at this stage, a motivated architectural hypothesis whose benefits the planned user study is meant to test.

The distinction also rests on a contested premise. By treating facial expression recognition as culturally neutral, VERS follows basic-emotion theory [5], whereas constructionist accounts and cross-cultural evidence suggest that part of the cultural variability already arises in how emotions are expressed and perceived [28, 30, 31, 6]. In VERS this variability is not modeled by the sensing stage; a natural extension is to make the perceptual stage itself culture-conditioned (e.g. cluster-specific calibration of the classifier), while keeping the inspectable separation of the two stages.

The use of macro-clusters is a necessary but reductive engineering simplification: it flattens national specificities (e.g. Japan vs. China within the same cluster). Importantly, the resulting error is *not uniform across colors*, nor across clusters: the Africa macro-cluster, which aggregates highly heterogeneous color-symbol systems, is the coarsest approximation and limits the generalizability of the chromatic matrix. Invariant color–emotion pairs can be handled safely, because the probability of cultural misunderstanding is low; high-variance associations (white, purple, yellow), by contrast, demand a much finer granularity than the current "West/East" partition. Ignoring this differential variance would expose the agent to actuation errors precisely in the most delicate cultural nuances. Moreover, the specific triples of Table 1 were authored manually and have not been checked for inter-rater agreement, which is a prerequisite for treating them as more than documented design hypotheses.

Although FER2025 is balanced for emotion classes, demographic representation may not be perfectly uniform. Beyond ethnicity, a residual gender bias may persist: features such as make-up or facial hair alter facial landmarks, so an unevenly exposed model could perform worse on one group than another, compromising the universality of the service. These demographic biases were not quantified in the present study, since per-group accuracy was not measured; doing so is a precondition for deployment.

The system currently assumes that the chosen color is effective, based on the color-psychology literature, without real-time physiological confirmation. Since indices such as Heart-Rate Variability (HRV) are established objective correlates of emotion regulation [40], the absence of biometric feedback

is a genuine limitation.

A first, essential direction is a large-scale cross-cultural user study: the A/B demonstration reported here is illustrative, and a systematic validation with human participants across the target cultural clusters, with statistical analysis against a culturally-neutral baseline and an inter-rater check of the chromatic matrix, is planned and constitutes the core of our future work. Four further directions follow. (1) *Micro-adaptive regulation*: replacing static macro-clusters with a granular cultural database or a reinforcement-learning agent that learns each user's chromatic preferences over time. (2) *Fairness and de-biasing*: stratified data collection balanced by gender and age, toward algorithmic fairness that keeps accuracy uniform across biological characteristics. (3) *Ambient perceptual calibration*: adapting emitted brightness and saturation to ambient light, since the same color has different psychophysiological impact in a dark room versus daylight. (4) *Closed biometric loop*: integrating remote photoplethysmography (rPPG) to estimate HRV from facial color variations, and pupillometry as a non-contact index of arousal and cognitive load [41], turning VERS into a genuinely closed-loop affective framework.

7. Conclusion

We presented VERS, a culturally-aware visual emotional regulation architecture for social robotics that separates universal perception from culture-specific signification. A systematic, data-centric study identified a Swin Transformer trained on a balanced large-scale corpus as the most reliable perceptual engine (87.87%), and a real-time prototype, in a preliminary proof of concept, illustrated culturally-differentiated chromatic regulation coupled with transparent inner-speech verbalization; its large-scale cross-cultural validation is the next step. Beyond its technical results, VERS is a concretely interdisciplinary artifact, joining computer vision, cross-cultural psychology and affective HRI, that reframes affective robotics as a problem of *meaning*, not only of recognition. Moving from macro-clusters to micro-adaptive, biometrically closed-loop regulation is, we argue, the path toward robots that are not merely accurate but culturally and empathically appropriate.

Declaration on Generative AI

During the preparation of this work, the author(s) used a generative AI assistant in order to: draft, revise and format the manuscript in the CEUR-WS style and perform grammar and spelling checks. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. Breazeal, Emotion and sociable humanoid robots, International journal of human-computer studies 59 (2003) 119–155.

[2] D. Feil-Seifer, M. J. Matarić, Socially assistive robotics, IEEE robotics & automation magazine 18 (2011) 24–31.

[3] M. Atari, M. Xue, P. Park, D. Blasi, J. Henrich, Which humans?, 2023.

[4] B. Berlin, P. Kay, Basic color terms: Their universality and evolution, Univ of California Press, 1991.

[5] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (1992) 169–200.

[6] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, S. D. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, Psychological Science in the Public Interest 20 (2019) 1–68. doi:10.1177/1529100619832930.

[7] A. J. Elliot, M. A. Maier, Color psychology: Effects of perceiving color on psychological functioning in humans, Annual review of psychology 65 (2014) 95–120.

[8] A. Gjaci, C. T. Recchiuto, A. Sgorbissa, Towards culture-aware co-speech gestures for social robots, International Journal of Social Robotics 14 (2022) 1493–1506.

[9] S. Šabanović, C. C. Bennett, H. R. Lee, Towards culturally robust robots: A critical social perspective on robotics and culture, in: Proc. HRI Workshop on Culture-Aware Robotics, 2014.
[10] F. M. Adams, C. E. Osgood, A cross-cultural study of the affective meanings of color, Journal of cross-cultural psychology 4 (1973) 135–156.
[11] D. Jonauskaite, A. Abu-Akel, N. Dael, D. Oberfeld, A. M. Abdel-Khalek, A. S. Al-Rasheed, J.-P. Antonietti, V. Bogushevskaya, A. Chamseddine, E. Chkonia, et al., Universal patterns in color-emotion associations are further shaped by linguistic and geographic proximity, Psychological science 31 (2020) 1245–1260.
[12] T. J. Madden, K. Hewett, M. S. Roth, Managing images in different cultures: A cross-national study of color meanings and preferences, Journal of international marketing 8 (2000) 90–107.
[13] R. W. Picard, Affective Computing, MIT Press, 1997.
[14] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al., A systematic review on affective computing: Emotion models, databases, and recent advances, Information Fusion 83 (2022) 19–52.
[15] A. Lindsay, A. A. Ramirez Duque, M. E. Foster, R. P. Petrick, Using ai planning for managing affective states in social robotics, in: Companion of the 2024 ACM/IEEE international conference on human-robot interaction, 2024, pp. 679–683.
[16] D. Löffler, N. Schmidt, R. Tscharn, Multimodal expression of artificial emotion in social robots using color, motion and sound, in: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 2018, pp. 334–343.
[17] A. Pörtner, L. Schröder, R. Rasch, D. Sprute, M. Hoffmann, M. König, The power of color: A study on the effective use of colored light in human-robot interaction, in: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, 2018, pp. 3395–3402.
[18] C. A. Thorstenson, Y. Tian, Dimensional approach for using color in social robot emotion communication, Color Research & Application 50 (2025) 424–432.
[19] A. S. Arora, A. Arora, J. R. McIntyre, Transforming customer experience in social robotics through explainable and interpretable artificial intelligence over a decade, Frontiers in Robotics and AI 13 (2026) 1693379.
[20] A. Chella, A. Pipitone, A. Morin, F. Racy, Developing self-awareness in robots via inner speech, Frontiers in Robotics and AI 7 (2020) 16.
[21] A. Pipitone, A. Chella, What robots want? hearing the inner voice of a robot, iScience 24 (2021) 102371.
[22] A. Pipitone, S. Corvaia, A. Chella, Towards robot affective appraisal linking inner speech and emotion, Robotics and Autonomous Systems (2026) 105363.
[23] J. X.-Y. Lek, J. Teo, Academic emotion classification using fer: A systematic review, Human Behavior and Emerging Technologies 2023 (2023) 9790005.
[24] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, A. C. Sobieranski, A survey on facial emotion recognition techniques: A state-of-the-art literature review, Information Sciences 582 (2022) 593–617.
[25] N. Park, S. Kim, How do vision transformers work?, arXiv preprint arXiv:2202.06709 (2022).
[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF international conference on computer vision (ICCV), Ieee, 2021, pp. 9992–10002.
[27] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, R. Seguier, Learning vision transformer with squeeze and excitation for facial expression recognition, arXiv preprint arXiv:2107.03107 (2021).
[28] L. F. Barrett, Are emotions natural kinds?, Perspectives on Psychological Science 1 (2006) 28–58. doi:10.1111/j.1745-6916.2006.00003.x.
[29] J. A. Russell, Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies, Psychological Bulletin 115 (1994) 102–141. doi:10.1037/0033-2909.115.1.102.
[30] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, P. G. Schyns, Facial expressions of emotion are not culturally universal, Proceedings of the National Academy of Sciences 109 (2012) 7241–7244.

doi:10.1073/pnas.1200155109.
[31] M. Gendron, D. Roberson, J. M. van der Vyver, L. F. Barrett, Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture, Emotion 14 (2014) 251–262. doi:10.1037/a0036052.
[32] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.
[33] P. Valdez, A. Mehrabian, Effects of color on emotions, Journal of Experimental Psychology: General 123 (1994) 394–409. doi:10.1037/0096-3445.123.4.394.
[34] S. Ronen, O. Shenkar, Mapping world cultures: Cluster formation, sources and implications, Journal of International Business Studies 44 (2013) 867–897.
[35] R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, V. Gupta, Culture, leadership, and organizations: The GLOBE study of 62 societies, Sage publications, 2004.
[36] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: A report on three machine learning contests, in: International conference on neural information processing, Springer, 2013, pp. 117–124.
[37] E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM international conference on multimodal interaction, 2016, pp. 279–283.
[38] P. Ekman, W. V. Friesen, A new pan-cultural facial expression of emotion, Motivation and emotion 10 (1986) 159–168.
[39] M. Adhavan, FER2025: Large-scale balanced facial expression dataset, Hugging Face Datasets, 2025. URL: https://huggingface.co/datasets/imadhavan/FER2025. doi:10.57967/hf/8296, license CC BY-NC 4.0.
[40] B. M. Appelhans, L. J. Luecken, Heart rate variability as an index of regulated emotional responding, Review of general psychology 10 (2006) 229–240.
[41] F. Guo, M. Li, Q. Qu, V. G. Duffy, The effect of a humanoid robot's emotional behaviors on users' emotional responses: Evidence from pupillometry and electroencephalography measures, International Journal of Human–Computer Interaction 35 (2019) 1947–1959.