From Perception to Social Behaviour: An Adaptive Engagement Pipeline for 3D Avatars*

Stefano Monfalcone[1], Gabriele Ragusa[1], Giovanna Runfola[1], Irene Maria Genovese[1], Agnese Augello[2], Silvia Franchini[2] and Luca Sabatucci[2,*]

[1]University of Palermo, Palermo, Italy
[2]Institute for High Performance Computing and Networking, National Research Council of Italy, Palermo, Italy

Abstract

Embodied conversational avatars are increasingly expected to behave in socially plausible ways, particularly in how they recognise participants and distribute attention during interaction. This paper presents a Unity-based adaptive multimodal engagement architecture for a 3D avatar. The system integrates AI-driven perceptual components for face analysis, head-pose estimation, and active-speaker detection within a decision-making framework that dynamically identifies the most relevant interlocutor and adapts the avatar's gaze accordingly. Its modular design allows perceptual components to be recomposed or differently weighted to support distinct social roles, as illustrated through two socially situated scenarios that foreground questions of participation, attention, and artificial agency in Digital Humanities contexts.

Keywords

embodied conversational avatars, multimodal engagement, gaze behaviour, social attention, human-avatar interaction, digital humanities

1. Introduction

Virtual environments and embodied conversational avatars are increasingly employed as museum guides, educational mediators, interactive storytellers, and participants in collaborative or performative experiences [1, 2]. The quality of the experience increasingly depends not only on visual realism, but also on how effectively avatars interpret human behaviour, recognise participation, and adapt their verbal and non-verbal responses in socially appropriate ways [3, 4]. This requires addressing an interactional question: how should an artificial presence distribute attention, acknowledge participants, and enact a socially recognisable role? Building a believable avatar therefore requires a balanced integration of realistic visual design, multimodal perception, and interactional coherence.

This paper focuss on the specific problem of recognising the participants involved in an interaction and determining how the avatar should distribute its attention according to its assigned social role and the interactional practice being supported. Building on recent work on multimodal conversational engagement for robot-to-human interaction, we adapt and customise the underlying perceptual components to the specific constraints of a real-time virtual avatar.

We present the design and Unity-based implementation of an adaptive multimodal engagement architecture for a 3D avatar, together with an evaluation agenda for its future empirical validation. The resulting system integrates face detection, head-pose estimation, and active-speaker detection. The perceptual data produced by these components is processed by a configurable engagement model and a behaviour tree that regulates the selection and persistence of the avatar's attentional target according to contextual social rules.

ReDH 2026: Rethinking the Future of Digital Humanities through Ethics and Artificial Consciousness, co-located with AIxIA 2026, October 6–9, 2026, Perugia, Italy

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.
stefano.monfalcone@community.unipa.it (S. Monfalcone); gabriele.ragusa01@community.unipa.it (G. Ragusa); giovanna.runfola@community.unipa.it (G. Runfola); irenemaria.genovese@community.unipa.it (I. M. Genovese); agnese.augello@icar.cnr.it (A. Augello); silvia.franchini@icar.cnr.it (S. Franchini); luca.sabatucci@icar.cnr.it (L. Sabatucci)



© 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

CEUR-WS.org/Vol-4314/ReDH3.pdf

A key feature of the proposed architecture is its adaptivity. In the current system, the social role of the avatar and the interactional practice are specified during the initial configuration. Adaptation then concerns the way in which the same perceptual information is interpreted through different variables, weights, and behaviour-tree conditions during the interaction. Behaviour trees were selected as the decision-making framework because they provide modularity, explainability, and deterministic execution, making them particularly suitable for socially aware avatar control.

We illustrate this form of adaptivity through two scenarios that differ in whom the participants address: 1) a listener scenario, where the avatar sustains its attention on a designated interlocutor; and 2) a meeting moderator, where the avatar distributes attention across several participants while considering accumulated speaking time, unsuccessful attempts to intervene, and under-participation. These scenarios are relevant to cultural, educational, and entertainment settings involving AI avatars, such as museum guidance, interactive learning experiences, and performances involving audience interaction.

The contribution of this work is therefore threefold. First, it presents the implementation of a multimodal perceptual layer within Unity and discusses the specific requirements of virtual embodiment. Second, it introduces a configurable engagement model and a behaviour-tree-based decision mechanism through which the same perceptual information can support different attention-management strategies. Third, it outlines an evaluation agenda that separates the technical assessment of perception and target selection from the subjective assessment of the resulting interaction.

2. Background and Related Works

Building social avatars involves a series of challenges that stem from different aspects of how social interactions function [5, 6]. From perception to decision making and finally action, various elements have to be taken into account for an avatar to perceive users, and interact with them.

A central concept in this research field is engagement, a multidimensional phenomenon involving attention, participation, affect, and commitment to the ongoing interaction [7, 8, 9]. A substantial part of the literature approaches engagement from the user's perspective, seeking to infer attention, participation, or involvement from observable multimodal cues. In this sense, engagement detection concerns the recognition of the user's current relation to the interaction. A complementary line of research concerns the behaviour of the artificial agent itself and how it should acknowledge, address, or maintain attention towards different participants. Recent reviews of embodied conversational agents emphasise that engagement detection and socially adaptive behaviour require the integration of verbal, visual, and contextual information [3]. From the perspective of an avatar, recognising a face in front of the camera or detecting speech alone is often insufficient to determine how a participant is contributing to the interaction or how the avatar should respond.

Gaze management is particularly relevant in multiparty interaction, where it contributes to signalling attention, identifying addressees, and regulating turn-taking: gaze, bodily behaviour, and speech jointly provide important information for understanding and predicting conversational dynamics in group settings [10]. Gaze direction can therefore be considered a socially meaningful action through which an avatar acknowledges a speaker, sustains attention on a selected participant, or redistributes attention across the group.

Beyond estimating engagement, another fundamental question concerns how a socially plausible digital person can be created [11]. To create a social avatar, both aesthetic and behavioural aspects, including verbal and non-verbal behaviour, must be considered. In particular, poorly coordinated gaze and body movements may contribute to the uncanny valley effect and cause discomfort in the people interacting with the avatar. Indeed, during a conversation, a social agent does not merely track a face, but participates in a social practice.

As highlighted in [12], agent models based mainly on mechanical reactions may fail to capture the social nature of human behaviour, relying on responses disconnected from the collective context. This observation motivates approaches that go beyond purely reactive strategies, such as continuously

following the most salient face or the current speaker, and instead take into account the social function assigned to the agent.

From a technical point of view, recent embodied conversational agents rely on multimodal perception pipelines that combine visual and audio cues to estimate users' attention and conversational engagement. Typical perceptual components include face detection for participant localisation, head-pose estimation for head-orientation and attention analysis, and active-speaker detection to associate speech with visible interlocutors. These components are generally implemented using deep-learning models that have achieved high levels of robustness and real-time performance, making them suitable building blocks for interactive applications.

Face detection is commonly addressed using CNN-based detectors, including RetinaFace [13], YOLO-based approaches [14], and UltraFace [15]. Head-pose estimation is commonly performed using deep regression networks [16], including architectures such as 6DRepNet [17], while audiovisual active-speaker detection is typically performed using multimodal models that jointly analyse facial dynamics and speech signals to determine which visible participant is speaking [18, 19, 20, 21]. These models provide the perceptual evidence needed to describe the interaction, but do not independently determine which participant should become the avatar's attentional target.

On the decision-making side, hierarchical and modular architectures, most notably behaviour trees, are widely adopted to control the behaviour of interactive and social agents, as they allow decision logic to be expressed through explicit, interpretable rules that can be configured according to the agent's role and the interactional scenario. This is particularly relevant when engagement values, however estimated, need to be translated into concrete actions, for example selecting whom to look at, how long to maintain the current target, when to disengage, or how to redistribute attention, a step that the perceptual pipeline alone does not address.

Against this background, the present work focuses on engagement from the agent-side perspective, considering multimodal and interactional cues to determine the relative priority with which the avatar should direct or maintain its attention towards each participant.

3. Engineering the Social Avatar

The proposed architecture provides a modular and adaptive framework in which multimodal information flows from perception to socially aware avatar behaviour. Figure 1 shows three core layers: a perceptual layer that extracts participant-specific multimodal information; an engagement model configured according to the avatar's social role and the interactional practice being supported, which estimates how the avatar should distribute its attention among the participants; and an attention tree that regulates the selection and maintenance of the attentional target. The resulting decision is passed to the Unity control system, which generates the corresponding gaze and body behaviour.

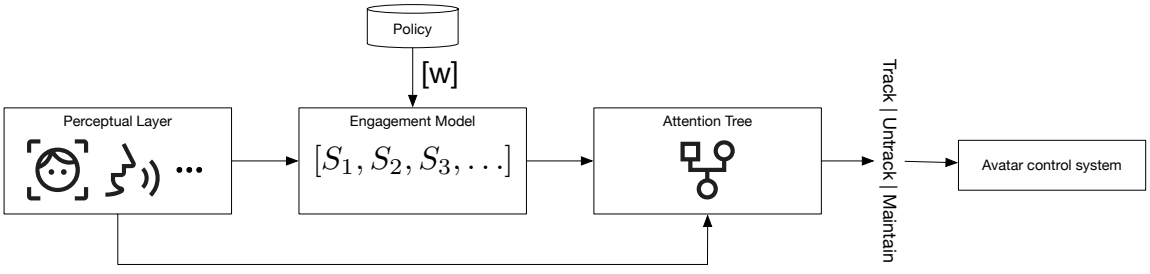


Figure 1: Overall Architecture of the Proposed Adaptive Engagement Pipeline. The proposed pipeline combines multimodal perception, adaptive engagement modelling, and behaviour-tree-based decision making to transform audiovisual observations into context-dependent socially-aware avatar behaviour.

3.1. Implementing the Avatar within Unity

Before examining the individual components, the architecture must be situated within the virtual environment in which it operates. To clarify the specific requirements of virtual embodiment, it is useful to compare the avatar with a physically embodied agent, such as a robot, in which gaze and body orientation are produced through mechanical actuation. Although both robotic and virtual agents can express engagement through these behaviours, the mechanisms used to generate them differ substantially.

In implementing the system, we must therefore consider the specific characteristics of Unity, our platform of choice, and how it supports the embodiment of the avatar's attentional behaviour. Unlike a robot, which relies on physical actuators, a virtual avatar must simulate movement through its skeletal structure and animation system.

This can be achieved using Unity's built-in Inverse Kinematics (IK) system [22], which allows the avatar's movements to be controlled without manipulating each skeletal joint individually. Rather than explicitly specifying the rotation of every bone, the system defines the target towards which the avatar should orient, while Unity automatically computes the corresponding joint rotations through its animation system. A single script attached to Unity's Animator component controls the relative IK weights assigned to the eyes, head, and upper body.

This approach is also useful when seeking to reduce behaviours that may contribute to the uncanny valley effect [11]. One possible strategy is to make the avatar dynamic instead of static, for example by simulating subtle human-like behaviour such as breathing and small movements of the limbs associated to it. Every movement must be carefully regulated so that the avatar does not appear excessively animated or too jerky when turning toward the engaged person, thereby supporting a more plausible and less disturbing interaction.

Another aspect to consider in achieving a natural interaction is the positioning of the virtual camera through which the user observes the avatar. We decided to place this camera approximately at the avatar's eye level, giving users the impression of interacting with the avatar through a video call. From the opposite perspective, the physical camera used by the avatar to perceive the users (typically a webcam positioned above the display) is equally important. Positioning this camera too far to the left or right, or significantly above or below the display, would create a mismatch between the users' actual positions and the directions towards which the avatar orients its gaze, potentially resulting in inaccurate gaze behaviour and user discomfort.

Another important difference between virtual avatars and robotic platforms lies in the way perceptual data is processed. In robotic platforms, AI models are typically developed and executed directly using deep-learning frameworks such as PyTorch. Unity applications, however, are developed in C#, whereas most frameworks used to define, train, and deploy neural networks are Python-based. As a result, a compatibility layer is required to integrate trained AI models into the Unity environment. Although several approaches have been proposed to enable interoperability between PyTorch and C#, a simpler and more widely adopted solution is provided by ONNX [23]. ONNX is an open-source framework that enables AI models to be deployed across different platforms while supporting hardware acceleration technologies such as CUDA. An additional advantage of ONNX is that the elementary operations performed by neural networks are optimised for efficient execution. Consequently, trained models must first be exported to the ONNX format using the export tools provided by the original framework (e.g., PyTorch [24]), so that they can be executed by the ONNX Runtime. Within Unity, the ONNX Runtime can be installed as a NuGet package using NuGetForUnity [25], a plugin that enables the use of NuGet packages within Unity projects.

3.2. Perceptual Layer

The perceptual layer receives the video stream captured by the webcam and the corresponding audio signal, and extracts participant-specific information concerning presence, spatial position, head orientation, and speaking activity. Its organisation builds on the multimodal perception approach proposed

by De Simone et al. [26], originally developed for robot-to-human interaction. In our work, these perceptual components are adapted to operate within Unity and to provide the information required by the subsequent engagement and decision mechanisms.

We designed a modular architecture for this layer intended for individual perceptual models to be replaced or extended without impacting other components.

The first component of the perceptual layer identifies the participants visible in each video frame. For this purpose, the system employs UltraFace [15], a lightweight convolutional neural network designed for real-time face detection. For each detected face, the model returns a bounding box that specifies its position and extent within the image. A tracking module then associates detections across consecutive frames, allowing the system to maintain a consistent representation of each participant and update the corresponding information over time.

Each detected face is subsequently processed by 6DRepNet [17], a deep neural model for unconstrained head-pose estimation. Given the image region containing a face, the model estimates the three-dimensional orientation of the participant's head with respect to the camera. This provides a non-verbal cue about the direction towards which the participant is oriented during the interaction.

Speaking activity is estimated using LR-ASD [27], a lightweight audiovisual neural network for active-speaker detection. The model receives a short sequence of face images associated with a participant together with the corresponding audio segment, and estimates whether that person is producing the detected speech. Speaking activity can therefore be attributed to a specific participant rather than inferred from the audio signal alone. The predictions obtained over the temporal sequence are aggregated into a participant-specific speaking estimate.

At each processing cycle, the perceptual layer provides, for every tracked participant, the position and centre of the face, the estimated head orientation, and an audiovisual estimate of speaking activity. These outputs describe the interaction as perceived by the system.

3.3. Engagement Model

To determine how the avatar should distribute its attention among the detected participants, the outputs of the perceptual layer are combined with temporal, spatial, and contextual information. Their contribution is regulated through a configurable set of weights ($[W]$ in Figure 1) reflecting the social function assigned to the avatar.

To formalise this process, each detected participant i is associated with an engagement score S_i, obtained by combining normalised cues derived from the perceptual outputs, the recent interaction history, and the active scenario configuration:

$$S_i = w_v V_i + w_a A_i + w_t T_i + w_g G_i + w_c C_i - w_p P_i + w_r R_i + w_u U_i, \qquad w_p \geq 0. \tag{1}$$

The quantities V_i, A_i, T_i, G_i, C_i, P_i, R_i, and U_i are input variables, whereas w_v, w_a, w_t, w_g, w_c, w_p, w_r, and w_u are their associated weights. Specifically, V_i and A_i are obtained from the visual and audio perceptual outputs and represent evidence of speaking activity. The spatial variable C_i is computed from the position of the detected face within the frame. The variables T_i, G_i, and U_i derive from the recent interaction history and represent, respectively, normalised accumulated speaking time, a normalised measure of unsuccessful attempts to intervene, and under-participation relative to the other members of the group. Finally, R_i is assigned by the scenario configuration and represents the role relevance of participant i.

The quantity $P_i \in [0, 1]$ is a non-negative perceptual penalty variable representing the overall reduction in confidence associated with assigning attention to participant i. Rather than describing a single physical quantity, P_i aggregates several participant-specific conditions that may reduce the reliability of the perceptual evidence available to the system. These conditions include, for example, increased distance from the camera, unstable face tracking, partial occlusions, poor illumination, low-confidence detections, degraded audio quality, or other adverse sensing conditions. Consequently, larger values of P_i correspond to lower confidence in the perceptual observations and therefore produce a greater reduction of the engagement score. Its contribution is written as $-w_p P_i$ because the effect of

increasing perceptual uncertainty should decrease, rather than increase, the engagement score. Since $w_p \geq 0$, the algebraic coefficient of P_i is $-w_p \leq 0$. Therefore, if P_i increases by an amount $\Delta P_i > 0$, its contribution changes by $-w_p \Delta P_i \leq 0$. The parameter w_p controls the strength of the perceptual penalty, whereas the minus sign determines its direction.

All input variables are normalised to the interval $[0, 1]$ before being combined. Raw scores are then normalised into a probability distribution over the detected participants through a softmax function:

$$Pr_i = \frac{e^{S_i}}{\sum_j e^{S_j}}. \tag{2}$$

The value Pr_i represents the normalised engagement priority assigned by the avatar to participant i relative to the other individuals detected in the scene. The resulting values satisfy

$$0 < Pr_i \leq 1, \qquad \sum_i Pr_i = 1,$$

with $Pr_i < 1$ whenever at least two participants are included in the normalisation.

Equation (1) is a configurable template, indeed, the weights can be chosen separately for each application scenario according to the intended behaviour of the avatar and the relative importance assigned to the corresponding cue (see Section 4).

3.4. Attention Tree

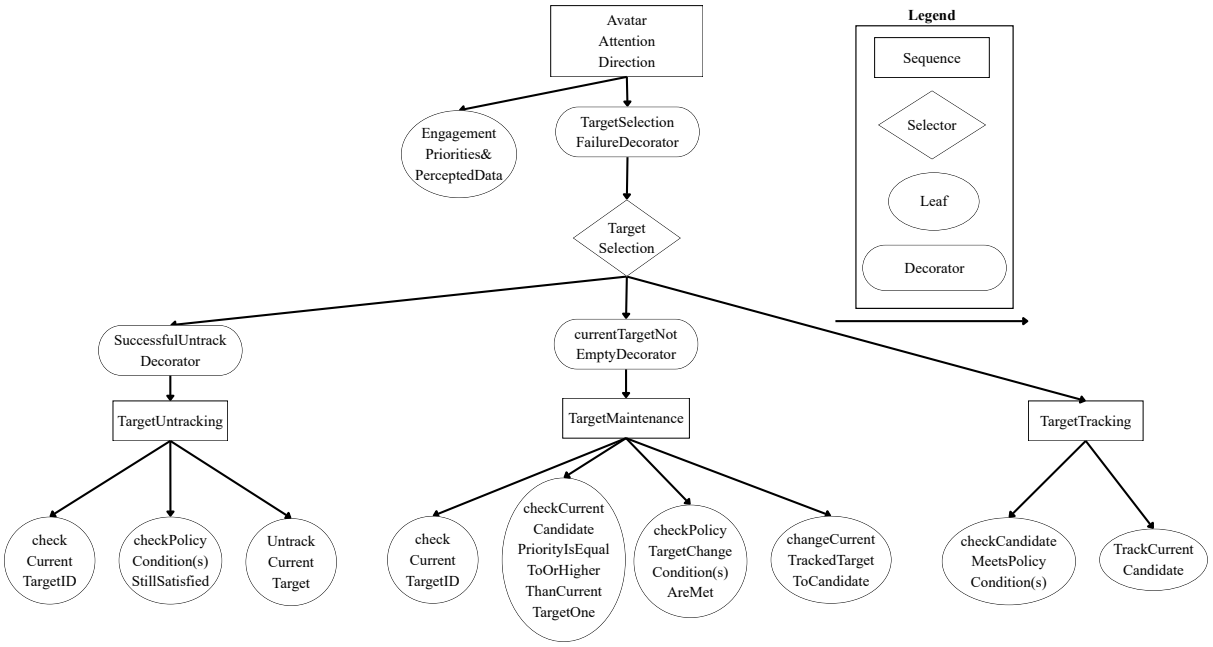


Figure 2: Architecture of the Attention Tree

The engagement values computed by the model do not directly trigger a gaze movement. Indeed, they are processed by a behaviour tree (namely the attention tree, see Figure 2), which determines whether and when modifying the current attentional behaviour.

Behaviour trees (BTs) are hierarchical control structures widely used to implement the decision-making logic of autonomous agents, particularly in videogames and robotics. A BT is evaluated from the root node at every execution cycle, traversing the tree until it identifies the behaviour to execute. Internal nodes define the control flow: sequence nodes execute their children in order until one fails, selector nodes evaluate alternative branches until one succeeds, and decorator nodes modify or constrain the execution of their child node. Leaf nodes implement the application-specific conditions and actions. To



Figure 3: A screenshot of the Social Avatar in Unity. The small inner picture is the image captured by the cam, and elaborated by the perceptual layer.

implement temporal persistence and long running behaviours, each node returns one of three execution states: Success, Failure, or Running. A node returning Running remains active and is re-evaluated at subsequent execution cycles until it eventually returns Success or Failure, allowing BTs to naturally represent movement, animation, or interaction while preserving a reactive decision-making process.

In our architecture, the engagement model and the attention tree perform complementary functions. The former estimates how the avatar's attention should be distributed among the participants, whereas the latter combines the relative priority assigned to each participant with the current state of the interaction and with perceptual and temporal conditions.

As shown in Figure 2, at each cycle, our attention tree first checks whether an attentional target is already active. If a target exists, the system verifies whether the corresponding participant is still detected and whether the conditions required to maintain the engagement remain satisfied. When the current target is no longer valid, the tree releases the attentional focus after the persistence interval has elapsed. If no target is active, the participant with the highest normalised engagement value is considered as a candidate. The candidate is selected only when the perceptual and operational conditions defined by the active policy are satisfied.

To implement this, the tree receives as input the perceptual data and the engagement priorities computed by the Engagement model, together with the avatar's current attentional state. First, the avatar checks whether the currently tracked user should be untracked by exploring the *TargetUntracking* branch. Regardless of the outcome, the *TargetMaintenance* branch is then explored. If a user is currently being tracked, the avatar checks whether the current target should be replaced by the current candidate. If the check succeeds, the target is updated; otherwise, the current target is maintained. In both cases, the tree exploration stops and returns the corresponding result. If no user is currently being tracked, the *TargetTracking* branch is explored to determine whether the current candidate should become the avatar's attentional target. If the check succeeds, the current target is set to the candidate and the tree exploration for that candidate terminates. Otherwise, the current target remains empty and the avatar stays in its neutral position (i.e., without directing its attention towards any user).

At the end, the attention tree can therefore produce three main outcomes: maintaining the current target, selecting a new participant, or releasing the attentional focus. Once a target has been selected, the centre of the corresponding face is passed to the Unity control system, which generates the associated gaze and body orientation.

It is worth noting that a temporal persistence mechanism prevents short interruptions in speaking activity or small fluctuations in the perceptual estimates from immediately causing a change of target.

The sourcecode of the presented architecture is available online[1], whereas Figure 3 shows a screenshot captured from Unity.

[1]https://github.com/ls-cnr/Unity3DAvatar

4. Application Scenarios

The engagement model and the attention tree can be configured to support different social functions without modifying the underlying perceptual components. The two scenarios considered in this work share the same perceptual layer and general decision architecture, but instantiate the engagement model by activating different variables, assigning them different weights, and applying different conditions for target selection and persistence.

The selected scenarios represent two recurring forms of interaction in which virtual agents may assume cultural, educational, or collaborative roles. A moderator, for example, must distribute attention across several participants and respond to imbalances in participation, whereas a listener must sustain attention towards a designated interlocutor despite the presence of competing signals. These practices are relevant to Digital Humanities settings in which avatars increasingly participate in guided interpretation, group discussion, learning, and other socially situated experiences.

4.1. Moderator Avatar

In the moderator scenario, the avatar is intended to support a balanced distribution of participation within a group, for example by acting as a co-tutor in an educational discussion. Its objective is therefore not simply to follow the person who is currently speaking, but to consider how participation is distributed over time.

The moderator configuration activates the variables associated with speaking activity, accumulated speaking time, unsuccessful attempts to intervene, and under-participation, while the remaining contributions of the general model are assigned a null weight. The resulting score is defined as

$$S_i^{\text{mod}} = w_v V_i + w_a A_i - |w_t|\, T_i + w_g G_i + w_u U_i, \tag{3}$$

where T_i is the normalised accumulated speaking time of participant i, G_i is a normalised measure of unsuccessful attempts to intervene, and $U_i \in [0, 1]$ measures under-participation relative to the other members of the group. In this configuration, $w_c = w_p = w_r = 0$; therefore, centrality, the perceptual penalty, and role relevance do not contribute to S_i^{mod}. The perceptual penalty P_i is intentionally disabled in the moderator configuration because the main objective is to promote balanced participation among all participants. Penalising participants for less favourable perceptual conditions, such as greater distance from the camera or less stable tracking, could further reduce the opportunity of less active participants to receive attention even when they show an intention to contribute. Perceptual validity is still enforced by the attention tree; setting $w_p = 0$ only prevents differences in perceptual reliability among valid participants from directly affecting the moderation score.

The term $-|w_t|T_i$ is the scenario-specific instantiation of the temporal contribution in Equation (1). In the moderator scenario, accumulated speaking time is used as a penalty because participants who have already spoken for longer should be progressively deprioritised. By contrast, G_i and U_i enter with positive contributions because unsuccessful attempts to intervene and limited participation increase the degree to which the avatar should direct its attention towards a participant. The magnitudes of the corresponding weights determine the strength of these effects. In particular, when $|w_t| > w_g$, an equal increase in the normalised temporal cue T_i has a larger effect in magnitude than an equal increase in the intervention cue G_i.

The participant with the highest normalised value is considered a candidate target. The attention tree then evaluates whether the current target should be maintained or whether attention should be redirected towards the new candidate. This distinction becomes particularly relevant when several conditions occur simultaneously, such as one participant dominating the conversation, another attempting to intervene, and a third contributing only marginally.

The resulting gaze behaviour can provide a non-verbal acknowledgement of participants who have had fewer opportunities to contribute. The avatar does not directly assign conversational turns, but its allocation of visual attention can support the moderation strategy encoded by the selected configuration.

4.2. Listener Avatar

In the listener scenario, the avatar is expected to maintain its attention on a predefined interlocutor rather than distribute it dynamically among all detected participants. A representative example is an avatar interacting with a particular student while other people remain present in the scene.

The listener configuration assigns a dominant weight to the role-relevance variable and sets the weights associated with group participation to zero. The role relevance of participant i is represented through a binary variable R_i, defined as $R_i = 1$ for the designated participant and $R_i = 0$ for all other individuals:

$$S_i^{\text{list}} = w_r R_i + w_v V_i + w_a A_i - w_p P_i, \tag{4}$$

where the manually configured weight w_r is assigned a larger value than the remaining active weights, while $w_t = w_g = w_c = w_u = 0$. The role-relevance term therefore gives the designated interlocutor a stable advantage, whereas V_i and A_i allow the current perceptual evidence to contribute to the score. This configuration allows the designated interlocutor to retain priority even when another participant produces a brief intervention.

In the listener configuration, P_i is retained because the objective differs from that of the moderator scenario: attention should remain primarily focused on the designated interlocutor rather than being distributed across all participants. In this configuration, P_i accounts for both the perceptual-reliability factors introduced in Eq. (1) and the competing perceptual evidence produced by non-target participants, so as to preserve attention towards the designated interlocutor. Its value is determined by the listener-specific decision policy using the available perceptual information, rather than being directly predicted as a separate output of the perceptual modules or learned from data. Larger values of P_i indicate that the contribution of participant i should be more strongly suppressed in order to preserve attention towards the designated interlocutor.

Since the algebraic coefficient of P_i is $-w_p \leq 0$, larger values of P_i reduce S_i^{list}. The manually configured parameter w_p determines the strength of this reduction, whereas the minus sign fixes its negative direction.

The designated participant therefore remains the preferred target while the corresponding perceptual and engagement conditions remain valid. The attention tree complements the weighted score through a persistence mechanism that maintains the current focus despite minor or short-lived interference. The target is released only when it is no longer perceptually valid or when the persistence conditions defined by the listener configuration are no longer satisfied.

5. Evaluation Agenda

A comprehensive evaluation of the proposed system requires examining both its technical performance and the quality of the resulting interaction. The evaluation agenda therefore combines objective measures of perception and attention management with subjective measures capturing how users experience the avatar's behaviour. Although such an evaluation has not been conducted within the present work, this section discusses the main methodological considerations and outlines the protocol that should guide the future empirical validation of the architecture.

An objective evaluation should address three complementary aspects: the performance of the perceptual components, the correctness and temporal stability of target selection, and the extent to which the moderator and listener configurations produce the intended attention-management behaviour.

Existing multimodal datasets may first be used to assess the individual perceptual components, including face detection, head-pose estimation, and active-speaker detection. However, the role-specific dynamics considered in this work—such as prolonged speaking, unsuccessful attempts to intervene, under-participation, target persistence, and secondary interventions—are not systematically represented in existing public datasets. A dedicated task-specific dataset should therefore be collected through controlled pilot sessions.

The collected interactions should be manually annotated with the active speaker, the duration and distribution of speaking turns, relevant attempts to intervene, and the participant who should receive the avatar's attention according to the configuration being evaluated. When the appropriate attentional target cannot be determined from a directly observable event, annotations should follow predefined scenario-specific guidelines and be compared across multiple annotators.

The perceptual components can be evaluated independently using standard measures such as accuracy, precision, recall, and F1-score. The complete decision process should then be assessed by comparing the target selected by the system with the annotated policy-consistent target. Relevant measures include target-selection accuracy, the frequency of inappropriate target changes, the duration of correct target maintenance, and the latency required to select or reacquire the appropriate participant.

The internal time series generated by the system, including avatar engagement scores, normalised engagement priorities, accumulated speaking time, and intervention-related variables, can also be analysed through descriptive and temporal statistics. These values may help characterise how the engagement model responds to changes in the interaction, but should not be treated as evidence of correctness in themselves. They should instead be interpreted in relation to the annotated reference data and the behaviour expected under each configuration.

Scenario-specific indicators are also required. In the moderator scenario, the analysis should consider speaking-time distribution, turn frequency, unsuccessful attempts to intervene, under-participation, and the frequency with which the avatar redirects its attention towards less active participants. These indicators can help determine whether the configuration supports a more balanced distribution of attention and participation. In the listener scenario, relevant measures include target-retention time, the number of inappropriate gaze shifts, and the time required to recover the designated interlocutor after a secondary intervention.

The role-specific configurations should also be compared with a non-adaptive baseline. A possible baseline may direct the avatar's attention towards the currently active speaker or apply a fixed weighting scheme independently of the avatar's role. Such a comparison would make it possible to determine whether the proposed configuration of engagement weights and decision conditions provides an observable advantage over a purely reactive attention strategy.

A user-perspective subjective evaluation should focus on perceived naturalness, perceived social attention, interaction comfort, and perceived engagement. Since these dimensions are closely related to the specific attention-management behaviours implemented by the avatar, it would be useful to develop and validate an ad hoc questionnaire tailored to the proposed interaction scenarios.

The first dimension, *perceived naturalness*, should assess the fluidity, realism, and contextual appropriateness of the avatar's gaze and body movements. The second, *perceived social attention*, should examine whether the avatar appears to follow the interaction, recognise the relevant participants, and direct or maintain its attention appropriately. The third, *interaction comfort*, should consider whether participants feel at ease and free to speak, without perceiving the avatar's gaze as intrusive, inappropriate, or disconnected from the ongoing interaction. A further dimension, *perceived engagement*, may assess whether the presence and behaviour of the avatar make the interaction more involving, interesting, or stimulating.

The questionnaire could use a five-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Each dimension should be assessed through multiple items, including a limited number of reverse-scored statements. General items may include: "The avatar's gaze movements appeared natural", "The avatar seemed aware of who was participating in the interaction", "The avatar directed its attention towards the appropriate person", and "I felt comfortable interacting in the presence of the avatar".

The instrument should also include items tailored to the interactional practice. In the moderator condition, participants may assess whether the avatar distributed its attention fairly across the group, acknowledged participants who had fewer opportunities to contribute, and avoided focusing excessively on dominant speakers. In the listener condition, participants may assess whether the avatar maintained appropriate attention towards the designated interlocutor and whether brief interventions by other participants caused distracting or unnecessary changes of focus.

Open-ended questions should complement the predefined items by allowing participants to report

moments in which the avatar's gaze appeared delayed, intrusive, unstable, or inconsistent with the ongoing interaction. These qualitative observations may reveal behaviours that are not fully captured by numerical ratings and may inform subsequent refinements of both the engagement model and the attention tree.

The questionnaire should first be administered in a pilot study to assess item clarity, relevance, and coverage of the intended constructs. Its validation should then examine the internal consistency of each dimension and, where the sample size permits, its factorial structure. Internal consistency may be assessed using Cronbach's alpha, while responses can be summarised through means, standard deviations, and response distributions. Scores may subsequently be compared across the adaptive and baseline conditions.

Combining objective system-level measures with subjective user feedback would provide complementary evidence regarding the technical reliability of the architecture, the effectiveness of its role-specific attention strategies, and the perceived quality of the resulting interaction.

6. Conclusions

This work has presented a modular engagement architecture for a 3D avatar that integrates a multimodal perceptual layer, a configurable engagement model, and a behaviour tree regulating target selection and persistence. In the current system, the social role of the avatar and the interactional practice are specified during the initial configuration. The corresponding variables, weights, and decision conditions are then used to adapt the avatar's attention behaviour throughout the interaction. The system does not currently interpret or estimate the social context autonomously.

The moderator and listener scenarios illustrate this form of adaptivity. The same perceptual information contributes differently to the avatar engagement score according to the configured social role, supporting either the redistribution of attention among participants or the maintenance of focus on a designated interlocutor.

Future work will validate the architecture through the evaluation agenda outlined in this paper, combining objective measures of perception and attention management with subjective user feedback.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] N. A. A. Halim, N. M. Suaib, F. E. Fadzli, Ai-powered avatar assistants: a systematic literature review of tools, application domains and challenges, Journal of Ambient Intelligence and Humanized Computing 17 (2026) 633–670.

[2] I. L. Garcia, E. Schott, M. Gohsen, V. Bernhard, B. Stein, B. Froehlich, Speaking with objects: Conversational agents' embodiment in virtual museums, in: 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), IEEE, 2024, pp. 279–288.

[3] F.-C. Yang, P. Acevedo, S. Guo, M. Choi, C. Mousas, Embodied conversational agents in extended reality: A systematic review, IEEE Access 13 (2025) 79805–79824.

[4] L. Gatto, G. Fulvio Gaglio, A. Augello, G. Caggianese, L. Gallo, M. La Cascia, Met-iquette: enabling virtual agents to have a social compliant behavior in the metaverse, in: 2022 16th International Conference on Signal-Image Technology Internet-Based Systems (SITIS), 2022, pp. 394–401.

[5] H. Ouhnni, A. Btissam, B. Meryam, K. El Bouchti, Z. Seghroucheni Yassine, S. N. Lagmiri, B. Rigalma, S. Ziti, The evolution of virtual identity: a systematic review of avatar customization technologies and their behavioral effects in vr environments, Frontiers in Virtual Reality 6 (2025) 1496128.

[6] B. Kaur, G. S. Josan, Bringing faces to life: A survey on realistic facial expressions in 3d virtual characters, International Journal of Human–Computer Interaction 42 (2026) 7906–7947.

[7] A. Pellet-Rostaing, R. Bertrand, A. Boudin, S. Rauzy, P. Blache, A multimodal approach for modeling engagement in conversation, Frontiers in Computer Science 5 (2023) 1062342.
[8] Q. Xu, L. Li, G. Wang, Designing engagement-aware agents for multiparty conversations, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 2233–2242.
[9] C. Yu, P. M. Aoki, A. Woodruff, Detecting user engagement in everyday conversations, arXiv preprint cs/0410027 (2004).
[10] M.-C. Lee, M. Trinh, Z. Deng, Multimodal turn analysis and prediction for multi-party conversations, in: Proceedings of the 25th International Conference on Multimodal Interaction, 2023, pp. 436–444.
[11] M. Musser, M. Mancheno, Building a digital person: Design best practices, https://www.fcatalyst.com/trends-and-signals/blog/aug2023/building-a-digital-person-design-best-practices, 2023.
[12] R. Mercuur, V. Dignum, C. M. Jonker, Integrating social practice theory in agent-based models: A review of theories and agents, IEEE Transactions on Computational Social Systems (2020).
[13] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5203–5212.
[14] D. Qi, W. Tan, Q. Yao, J. Liu, Yolo5face: Why reinventing a face detector, in: European Conference on Computer Vision, Springer, 2022, pp. 228–244.
[15] Linzaer, Ultra-light-fast-generic-face-detector-1mb, available online, last accessed: 7/24/2026. URL: https://github.com/linzaer/ultra-light-fast-generic-face-detector-1mb.
[16] A. Asperti, D. Filippini, Deep learning for head pose estimation: A survey, SN Computer Science 4 (2023) 349.
[17] T. Hempel, A. A. Abdelrahman, A. Al-Hamadi, 6d rotation representation for unconstrained head pose estimation, in: 2022 IEEE International Conference on image processing (ICIP), IEEE, 2022, pp. 2496–2500.
[18] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, L. Chen, A light weight model for active speaker detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22932–22941.
[19] J. L. Alcázar, F. Caba, A. K. Thabet, B. Ghanem, Maas: Multi-modal assignation for active speaker detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 265–274.
[20] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, et al., Ava active speaker: An audio-visual dataset for active speaker detection, in: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2020, pp. 4492–4496.
[21] Y. J. Kim, H.-S. Heo, S. Choe, S.-W. Chung, Y. Kwon, B.-J. Lee, Y. Kwon, J. S. Chung, Look who's talking: Active speaker detection in the wild, arXiv preprint arXiv:2108.07640 (2021).
[22] B. Siciliano, L. Sciavicco, L. Villani, G. Oriolo, Robotics: Modelling, Planning and Control, Springer Science & Business Media, 2010.
[23] ONNX: Open Neural Network Exchange, https://github.com/onnx/onnx, 2017.
[24] Pytorch to onnx exporter documentation, https://docs.pytorch.org/docs/2.9/onnx_export.html, 2026.
[25] NuGetForUnity: A NuGet Package Manager for Unity, https://github.com/GlitchEnzo/NuGetForUnity, 2019.
[26] G. De Simone, L. Greco, A. Saggese, M. Vento, Robot to human interaction with multi-modal conversational engagement, in: 2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR), IEEE, 2025, pp. 1–8.
[27] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, L. Chen, Y. Chen, Lr-asd: Lightweight and robust network for active speaker detection, International Journal of Computer Vision 133 (2025) 4749–4769.