

Cadence-Stratified Multimodal Perception for Real-Time Child–Robot Interaction

Marwan Mohammed^{1,2} and Javad Zarrin² and Salma Elsayed² and Vera Kempe¹

¹Faculty of Social and Applied Sciences, Abertay University, Dundee, UK

²Faculty of Design, Informatics and Business, Abertay University, Dundee, UK

m.mohammed@abertay.ac.uk

Abstract

Child–robot interaction places real demands on perception systems: a robot engaged in play with a child cannot afford to miss the child’s focus of attention, yet running every perception module at every frame is computationally expensive on the kind of edge hardware that privacy-respecting robots must use. This paper addresses this tension by proposing a cadence-stratified multimodal perception pipeline, WordMate-CRI, in which each module runs at a rate that trades computation against staleness, informed by how long each cue remains stable for a potential interaction partner. Detection and tracking run continuously; action recognition uses rolling temporal buffers; emotion is refreshed periodically; and attributes such as age and gender are estimated once at the start of a session and cached. We evaluated WordMate-CRI through an LLM response-quality ablation study across six toy-play scenarios and a latency benchmark on Jetson AGX Orin hardware. The ablation shows that VLM enrichment and object evidence are the dominant grounding signals, with WordMate-CRI achieving a mean scene-grounded score of 0.967. On live Jetson input, cadence-based scheduling raises throughput from 4.3 to 18.2 frames per second (FPS), clearing the 15 FPS deployment target, while reducing WordMate-CRI latency from 6411.5 to 5844.5 ms. Taken together, the results make a practical case that deciding when each perception module should run matters just as much as deciding what it should recognise. This optimisation is important for systems aiming to produce contingent and timely child-directed speech in multi-modal contexts.

1 Introduction

Child–Robot Interaction (CRI) has become an important area in social, educational, therapeutic, and assistive robotics. In play, learning, communication support, healthcare, and autism-related interventions, robots are increasingly expected to act as

responsive social partners for children (Scassellati et al., 2012; Kozima et al., 2007, 2009; Nadeem et al., 2025; Belpaeme et al., 2012; Efthymiou et al., 2022). Our project aims to design a robot capable of producing appropriate child-directed speech. This requires processing the attentional focus of the child in a timely and socially meaningful way, using multi-modal cues such as object detection, tracking of the child’s position and movement, action, affect, engagement, and stable person-level attributes (Duncan et al., 2024; Cooper et al., 2026; Faria et al., 2025; Efthymiou et al., 2022).

A central challenge in real-time CRI is not only determining what information should be perceived, but when each perceptual process should run. Existing multimodal CRI and HRI systems combine vision, affective computing, behavioural analysis, and contextual reasoning (Duncan et al., 2024; Zhao et al., 2025; Belpaeme et al., 2012; Efthymiou et al., 2022). However, the temporal execution of these modules is often treated as an implementation detail rather than as an explicit design problem. Here we propose a more principled solution, recognising that multi-modal cues do not change at the same rate: object handling, position, and movement may change at faster rates than affect expression, which may evolve over a longer interval, while attributes such as age and gender remain more stable even in multi-child play sessions.

This temporal mismatch creates a practical cost. If every module is executed at every frame, the computation and latency increase, especially when several deep-learning models are used on the same robot. Recent HRI benchmarking also shows that current vision-language perception models still face difficult accuracy–latency trade-offs for real-time deployment (Shi et al., 2025). At the same time, reducing the update rate without a clear policy may cause the robot to miss relevant changes in the interaction. This paper therefore treats temporal scheduling as part of the perception architecture

itself, rather than as a later optimisation step.

Our motivation is also consistent with evidence from psychology and neuroscience. Human perception is organised across multiple temporal scales: early sensory areas respond to rapidly changing input, while higher-level regions integrate information over longer windows (Hasson et al., 2008). A similar temporal structure has been shown during the comprehension of naturalistic stories (Lerner et al., 2011), event segmentation (Kurby and Zacks, 2008), adaptive temporal integration (Akyürek, 2025), and intrinsic neural timescales (Manea et al., 2022). These findings suggest a useful design principle for CRI: perceptual updates should be matched to the temporal stability of the to-be-processed cue.

Here we propose WordMate-CRI as a cadence-stratified multimodal perception architecture for real-time CRI. WordMate-CRI combines continuous detection and tracking, rolling-window action inference, periodic facial analysis, and cached demographic estimates. Section 3 specifies the update rules and their rationale.

Related work provides important foundations but does not directly address this specific scheduling problem. Multimodal HRI research highlights the need to coordinate heterogeneous perceptual channels in social environments (Duncan et al., 2024; Zhao et al., 2025). FlowAct shows the value of continuous perception and modular action subsystems (Dhaussy et al., 2024), while relevance-driven scheduling reduces latency by activating perception modules according to scene context (Huang et al., 2026). However, these approaches mainly schedule modules according to the relevance of the current scene. By contrast, the WordMate-CRI architecture assigns update cadences proactively, based on the expected temporal stability of the cues. Multiple-timescale and cognitive HRI models also show that robot processing need not be uniform in time (Yamashita and Tani, 2008; Umbrico et al., 2023, 2022), but their focus is mainly on reasoning, control, or cognitive organisation rather than module-level perception scheduling. The gap addressed in this paper is therefore a cue-stability-based temporal scheduling policy for multimodal child perception, where each module is updated according to the expected dynamics of the cue it estimates. Our main contributions are:

1. A cadence-stratified perception architecture of WordMate-CRI for real-time CRI, where

perception modules run at different update rates instead of being executed uniformly at every frame.

2. An implemented multimodal perception pipeline combining continuous child detection and tracking, rolling-window action recognition, periodic emotion analysis, and sparse estimation of stable attributes.
3. A theory-to-engineering link between evidence for multi-scale temporal processing and a practical scheduling design for CRI.

This work is guided by two research questions: (RQ1) How does cadence-stratified scheduling affect visual-processing throughput and latency relative to per-frame processing? (RQ2) How does the availability of each nonverbal cue, namely gaze and head orientation for attention, affect for response tone, and object and action evidence for grounding, affect the quality of the generated response? Object detection and the VLM enrichment step that grounds response content operate at the segment level, evaluated once every five seconds, while the cadenced perception modules operate at the frame level; the throughput and latency gains reported in Section 5.2, addressing RQ1, arise from reducing frame-level computation, while response quality depends most on the segment-level evidence examined in Section 5.1, addressing RQ2. These are two distinct parts of the pipeline rather than a single measured trade-off curve. Thus, we evaluate processing throughput and latency against a per-frame baseline and separately examine how contextual information affects the generated response.

2 Related Work

The design draws on multi-timescale models of perception, real-time multimodal HRI, and child-centred robot perception.

A recurring finding in cognitive neuroscience is that the cortical areas integrate information over different time windows, with early sensory regions tracking fast input and higher-level regions accumulating evidence over longer periods (Hasson et al., 2008). Related patterns have been observed in story comprehension (Lerner et al., 2011), event segmentation (Kurby and Zacks, 2008), adaptive temporal integration (Akyürek, 2025), and intrinsic neural timescales (Manea et al., 2022). These findings suggest that different perceptual cues are naturally handled at different temporal resolutions,

a principle that has begun to surface in robotics as well. Yamashita and Tani demonstrated that functional hierarchy can emerge from a multi-timescale neural network in a humanoid robot setting (Yamashita and Tani, 2008), and related HRI architectures separate fast reactive processing from slower deliberative reasoning (Umbrico et al., 2023, 2022). Layered systems such as LAAIR extend this further by combining reactive and deliberative components within a single autonomous interactive robot (Jiang et al., 2018). What these works share is an acknowledgement that processing need not be temporally uniform. What they do not address is how individual social-perception modules in CRI should be scheduled, which is precisely the gap this paper targets.

On the multimodal perception side, the challenge of coordinating heterogeneous perceptual channels under real-world constraints is well documented (Duncan et al., 2024; Zhao et al., 2025). Systems such as FlowAct have moved away from strictly sequential pipelines by introducing a continuous asynchronous loop between perception and action (Dhaussy et al., 2024), and scheduling work has shown that activating perception modules according to scene relevance can significantly reduce latency (Huang et al., 2026). The tension between precision and latency is particularly sharp in edge hardware: studies of open-vocabulary mobile robot perception (Park et al., 2025) and the HRI-Bench evaluation of vision-language models (Shi et al., 2025) both confirm that current systems still struggle to meet the performance and timing demands of real-time deployment simultaneously. These studies make a strong case for principled scheduling, but their focus stays on general HRI, human–robot collaboration, or mobile robotics. Child-centred perception, with its specific cue profile and interaction demands, has so far received little direct attention. The two directly comparable systems differ from the approach proposed here in when the update rate is decided. The asynchronous perception-action system cited above keeps environment state continuously current and coordinates modular action subsystems through a planner, without assigning perception modules a fixed update rate in advance. The relevance-driven scheduler cited above activates modules based on prior outputs and scene context, which is a reactive, relevance-triggered policy rather than a cadence set proactively from the expected stability of each cue. The schedule proposed in this paper is therefore distinguished

by assigning per-cue update rates ahead of runtime, evaluated specifically in a child-perception setting rather than in general human-robot interaction. That gap is only partly filled by the CRI literature. Early work established the therapeutic value of social robots for children, including simple platforms such as Keepon in autism-related settings (Scassellati et al., 2012; Kozima et al., 2007, 2009), and recent reviews confirm growing interest in robots that support communication, learning, and social behaviour in children with autism spectrum disorder (Nadeem et al., 2025). More recently, multimodal CRI has increasingly been framed as a socially situated process that requires coordinated perception and timely response (Belpaeme et al., 2012). For example, ChildBot demonstrated multi-robot perception and interaction with children at scale (Efthymiou et al., 2022); affective computing and biophysical signals have been integrated to make interactions emotionally aware (Faria et al., 2025), and incremental visual learning has been explored to handle child-specific appearance variation (Efthymiou et al., 2021). Engagement detection ties many of these threads together, drawing on gaze, head orientation, action, and affect simultaneously (Lu et al., 2024; Kim et al., 2023). However, what this body of work focuses on is *what* should be recognized and *how* recognition should be performed. The question of *when* each perceptual module should run during live interaction has received far less attention. This is particularly important for systems that aim to produce child-directed speech contingent on children’s focus of attention, thereby supporting word learning (Yu et al., 2019; Schroer and Yu, 2023). Rather than proposing new recognition models or a new multimodal CRI platform, we explore a cue-stability-based temporal scheduling policy: continuous updates for detection and tracking, rolling-window inference for behavioural cues, periodic refresh for affect, and sparse session-level estimation for stable attributes such as age and gender.

3 Proposed Pipeline

3.1 WordMate-CRI Overview

In the proposed cadence-stratified multimodal perception architecture each perception module operates at an update rate matched to the temporal stability of each cue. The system processes a continuous RGB-D video stream and produces temporally aggregated scene representations that drive

spoken language responses generated by a locally deployed LLM. The WordMate-CRI is structured across three computational tiers operating at distinct temporal granularities (Figure 1):

1. Tier A—Frame loop: person and object detection (YOLO) and multi-person tracking (Norfair) run on every processed frame. Action recognition (R(2+1)D-18) uses a rolling temporal buffer with stride-based inference. Head orientation (6DRepNet) is computed in the frame loop at its configured stride.
2. Tier B-Cadenced: asynchronous background workers whose update rate trades computation against cue staleness rather than assuming the underlying cue changes slowly. Facial affect and gaze (OpenFace3) refresh every 5 person-detection frames (every frame when head-pose is enabled), and child age and gender (MiVOLO) are polled every 8 person-detection frames and locked after 4 successful inference runs; both run as background workers whose results are collected within the same frame call.
3. Tier C-Segment-level (5-second windows): Temporal aggregation of Tier A and B observations into a structured scene record, parallel VLM enrichment on the segment keyframe and spoken response generation by the LLM.

Detection and tracking are updated continuously to follow object and person movement, while action inference integrates evidence over a temporal window. Facial analysis is refreshed periodically to reduce repeated computation, and demographic estimates are cached while the tracked person remains unchanged. Gaze-to-object association is updated each frame using the available estimates. Although the transitions in affect, gaze, and head pose are fast, the relevant cue values persist considerably longer: spontaneous facial expressions typically last 0.5 to 4 s, and looks to objects during toy play extend over seconds (Yu et al., 2019). At the deployed refresh interval of approximately 275 ms, each such event is sampled at least twice before it ends.

A natural question is whether the modular pipeline could be replaced by a single locally deployed VLM. This approach was deliberately avoided for four reasons: (1) privacy constraints preclude cloud-based VLMs since transmitting

children’s video violates GDPR and COPPA and diminishes user trust; (2) accuracy limitations of lightweight local VLMs (e.g., LLaVA-1.5-7B, Phi-3-Vision) make them unreliable for child toy-play content, where objects and actions are under-represented in standard training corpora; (3) latency incompatibility, as 7B+ parameter models require 500 ms–2 s per inference on Jetson AGX Orin, yielding frame rates below our adopted 15 FPS visual-processing target (see Section 4.3.3); and (4) hallucination rates in edge-deployed VLMs are difficult to calibrate, unlike the modular classifiers which offer bounded, characterisable uncertainty. The VLM is therefore retained solely as a segment-level scene-enrichment oracle.

3.2 Segment Aggregation and LLM Response Generation

At the segment level (Tier C), outputs from the cadenced perception modules are aggregated over a 5-second window into a structured context record. The aggregation is signal-specific: actions are selected by majority vote from the R(2+1)D-18 rolling-buffer classifier, with an action stability score used to flag possible activity switching when the dominant label covers less than 50% of the window. Emotion is summarised using the dominant OpenFace3 label, an expression distribution, and an optional trajectory field when the dominant emotion changes between the first and second half of the segment. YOLO object detections are deduplicated by label, ranked by average confidence, and filtered using a 30% presence-ratio threshold to remove transient detections. Gaze behaviour is represented through a focus ratio and the number of unique gaze targets, allowing the system to distinguish sustained attention from scattered exploration. Age and gender estimates from MiVOLO are aggregated over four successful inference runs, scheduled every eight person-detection frames, and then cached until the demographic lock is reset.

The resulting context record is used to generate the robot’s spoken response. A VLM (Gemma3-4B via Ollama) runs in parallel on the segment keyframe to provide high-level toy and activity hints, while a locally deployed LLM (Qwen2.5-3B-Instruct via Ollama) produces a four-line caregiver-style response: (1) a warm observation, (2) a gentle invitation, (3) a playful question, and (4) an encouraging follow-up. The final text is converted to speech using Piper TTS and delivered by the robot.

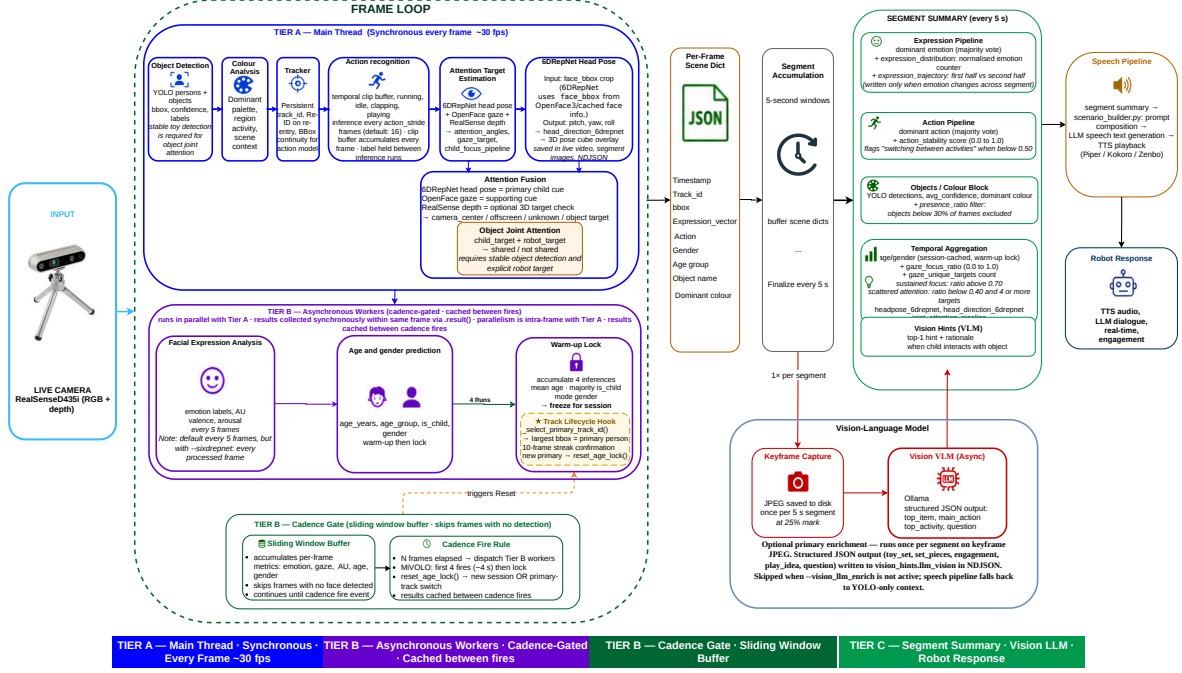


Figure 1: WordMate-CRI. Tier A (main thread, every frame) performs person detection, multi-person tracking, action recognition (rolling buffer with a 16-frame stride in the reported deployment benchmark), and head orientation estimation. Tier B (asynchronous background workers) refreshes facial affect and gaze every 5 person-detection frames and estimates age and gender every 8 person-detection frames, locking the latter after 4 successful inference runs. Tier C aggregates Tier A and B outputs once per 5-second segment, triggering VLM scene enrichment and spoken response generation.

4 Materials and Experimental Setup

4.1 Hardware

The pipeline ran on an NVIDIA Jetson AGX Orin 64GB Developer Kit as the main onboard compute unit. Visual input came from an Intel RealSense D435i RGB-D camera, providing aligned colour and depth streams for scene understanding and gaze-to-object association. Responses were played through a USB stereo speaker. Offline development and evaluation used a Windows workstation with an NVIDIA RTX 5000 Ada Generation Laptop GPU.

4.2 Software and AI Framework

We implemented the pipeline in an isolated Conda environment to support reproducible execution across both platforms. Visual perception combined YOLO object detection, Norfair tracking, R(2+1)D-18 action recognition, OpenFace3 facial analysis, 6DRepNet head-pose estimation, and a custom RGB-D gaze-to-object association module. MiVOLO provided demographic estimates after a short warm-up period. Responses were generated via Qwen2.5-3B-Instruct through Ollama, vision-

language enrichment by Gemma3-4B, and spoken output synthesised using Piper TTS.

4.3 Evaluation Design

4.3.1 Dataset and Settings

The response evaluation uses six proxy toy-play clips depicting block building, cup stacking, picture-book interaction, toy-kitchen play, doctor-kit play, and racetrack play. Each clip is 10 seconds long at 30 FPS and 640×480 resolution, giving 60 seconds of footage. Two approximately five-second windows per clip yield 12 scene segments. Evaluating each segment under eight named conditions produces 96 response records. Child demographic context is assigned for response generation and is not a measurement of the performers' ages. Recorded benchmarking processes saved video, whereas live benchmarking processes camera input on the Jetson AGX Orin and Windows workstation. The live evaluation contains five segment records per Jetson condition and three baseline/four proposed records on Windows. These counts describe timing records, not independent participants or matched repetitions of the same scene.

4.3.2 Ablation Study

The response study examines how changes to the available perceptual context affect scene-grounded caregiver-style language. Each generated response is evaluated against the same human-reviewed ground-truth reference for its scene segment across all conditions. This ground-truth reference comprises a verified scene description, an example of a contextually appropriate child-directed response for that scene, and scene keywords.

Ground-Truth: Ground truth was created from the output generated by running the WordMate-CRI across all six recorded clips. For each toy segment, saved frames, segment NDJSON files, VLM hints, generated responses, and prompt/response tables were manually reviewed. The ground-truth JSON was then refined where needed so that the scene description, reference caregiver response, and keywords reflected what was visible in the scene. This was necessary because automatic detections sometimes produced noisy labels, such as identifying colourful blocks as a skateboard or a bottle, which were not present in the scene. The human-reviewed ground truth was therefore used as the verified scene reference. It included a caregiver-style reference response, written as a warm observation, gentle invitation, playful question, and encouraging follow-up, together with scene keywords such as the toy object, action, and play activity. The ground-truth reference for each segment was initialised automatically from the recorded scene output and then manually reviewed against the corresponding video, saved frames, and VLM hints. Fields that visibly conflicted with the clip were corrected. These human-reviewed references were then used to evaluate the system’s generated responses.

Ablation Conditions: Each segment was evaluated under the WordMate-CRI and seven ablated conditions. WordMate-CRI retained all available segment information and served as the baseline. There are seven ablation conditions: no action recognition (-Action), no emotion detection (-Emotion), no gaze and focus-target information (-Gaze), no age and gender (-Age/Gender), no VLM scene hints (-Vision LLM), no object detections (-Object), and retention of solely object detections (Obj. only). These conditions were used to test how each signal affected robot response based on scene recognition.

Five of the seven conditions, namely no action

recognition, no emotion detection, no gaze and focus-target information, no age and gender, and no object detections, remove one perceptual signal at a time in order to isolate its marginal contribution to the generated response, following a leave-one-out design. The remaining two conditions test the extremes of visual-semantic grounding, directly motivated by the design question raised in Section 3.1 of whether a modular pipeline is preferable to interpretation based solely on the vision-language model: the no-VLM-scene-hints condition removes that enrichment step entirely, while the objects-only condition retains solely object-detection evidence. Because VLM scene hints are themselves partly derived from detected objects, the no-object-detections condition removes both the object list and the associated VLM hints together, and is therefore best read as a combined removal of object-derived evidence rather than an isolated test of object detection alone.

Evaluation Metrics: Each condition was evaluated using four complementary metrics. The Scene Grounded Score (SGS) was used as the primary measure of response quality and scene faithfulness, while SGS degradation, hallucination rate, and Keyword Hit Rate (KHR) were used to examine grounding loss, error patterns, and lexical overlap. SGS is an automated score, normalised between zero and one, computed from five weighted components, namely object grounding, action grounding, scene consistency, developmental-language appropriateness, and caregiver-response format, each contributing up to two points, from which a hallucination penalty of up to two points and a generic-language penalty of up to one point are subtracted; it reflects the alignment between a generated response and the verified scene and response format, and is not a validated measure of child language development. KHR is a particularly important parameter that ensures correct labelling responses by the robot to promote children’s word learning, as it is the proportion of reference scene keywords, matched through lemma-level overlap, that appear in the generated response, a lexical-coverage measure rather than a semantic one, reported alongside SGS rather than in its place. Correct object and action labelling in caregiver speech is the mechanism by which KHR connects to word learning (Yu et al., 2019; Schroer and Yu, 2023).

The ablation study in this section was run with the implementation’s default action-recognition stride of 8 frames, distinct from the stride of 16

Metric	Description
Per-frame processing time	Defined as the mean duration of <code>analyze_frame()</code> over a segment, measured by <code>T1_mean_frame_ms</code> . It indicates the frame-level cost of visual processing.
Frames per second (FPS)	Calculated as $1000/T1_mean_frame_ms$. It measures visual-processing throughput and is used to assess performance against the 15 FPS Jetson live target.
Vision processing time	Defined as $vision_proc_ms = T2 - T0$, measuring the wall-clock time from the first frame of the segment to segment finalisation.
Speech pipeline latency	Defined as $pipeline_ms = T7 - T_{seg_read}$, measuring the delay from segment pickup to the first spoken word.
System latency	Defined as $system_latency_ms = vision_proc_ms + pipeline_ms$. It summarises visual processing and speech-generation latency while excluding robot turn-taking time.

Table 1: Evaluation metrics used in Latency analysis.

frames used in the latency benchmark of Section 5.2; both reflect the same rolling-buffer mechanism at different sampling rates.

4.3.3 Latency Benchmark: Per-Frame Baseline versus WordMate-CRI

The timing study compares the per-frame baseline with WordMate-CRI using the recorded and live input settings described above. Table 1 defines the five reported timing measures. We deployed 15 FPS as a practical target for live visual-processing throughput on the Jetson, informed by (Park et al., 2025), who identified rates above 15 FPS as suitable for balancing responsiveness and perceptual capability in general HRI. This is an engineering benchmark for our deployment, rather than a validated minimum for child–robot interaction.

5 Results and Discussion

The ablation study showed that VLM scene enrichment and object evidence contributed most strongly to scene-grounded response quality (Section 5.1). Alongside these response-quality findings, the latency benchmark showed that cadence-based scheduling improved visual-processing throughput and reduced system latency, enabling live Jetson deployment to exceed the adopted 15 FPS target (Section 5.2). However, because response quality was not directly compared between per-frame and cadenced processing, these results do not establish accuracy preservation.

5.1 Ablation Study Results

The SGS results in Figure 2 show that WordMate-CRI (Full) consistently achieved strong scene-grounded response quality across all six toy scenarios, with a mean SGS of 0.967 across the twelve

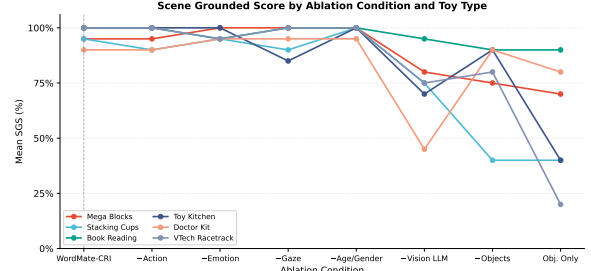


Figure 2: Scene Grounded Score across ablation conditions for each toy type.

evaluated segments. Removing action, emotion, gaze, or demographic information produced only minor changes, with mean SGS values of 0.958, 0.967, 0.950, and 0.992 respectively, suggesting these modalities had limited effect on core scene recognition, though they may still shape response detail. Larger reductions appeared when visual-semantic grounding was weakened: removing VLM context dropped the mean SGS to 0.733, removing object detections to 0.775, and the objects-only condition produced the lowest score at 0.567. This confirms that high-level visual interpretation and object evidence are complementary — object labels contribute useful grounding, but are not sufficient on their own for a contextually appropriate robot response.

The hallucination-rate Figure 3 results help explain the SGS reductions observed in the more severe ablations. When scene grounding was weakened, the model did not simply become less specific; in several cases it shifted towards an entirely different but plausible play scenario. In the Mega Blocks clips, for instance, the verified activity was block stacking, yet some ablated responses referred to a skateboard or bottle, likely because the colourful block structure was mislabelled by lower-level detectors. Without VLM hints to correct this, the model followed the noisy cues. Similar drift was observed across other toys: stacking cups responses moved towards ball play, toy kitchen clips occasionally shifted to ball rolling, and racetrack clips drifted towards cups, bubbles, or cooking-related language. Although these responses remained plausible within a child-play setting, they were not faithful to the verified scene.

The KHR shown in Figure 4 supports the SGS findings but also reveals why lexical overlap alone is an unreliable evaluation measure. The full pipeline sometimes received a lower KHR than an ablated condition because KHR rewards only

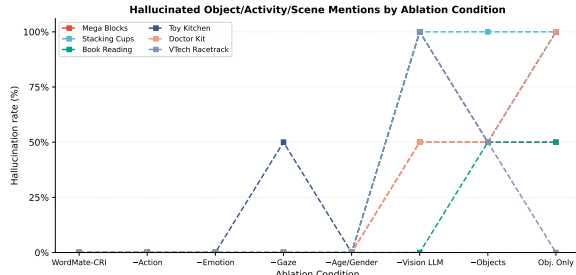


Figure 3: Rate of hallucinated object, activity, or scene mentions by ablation condition and toy type. Each toy-condition combination is computed from two evaluated responses, so plotted values fall at 0%, 50%, or 100%; they reflect rule-based hallucination flags rather than a fine-grained or exhaustively human-annotated rate.

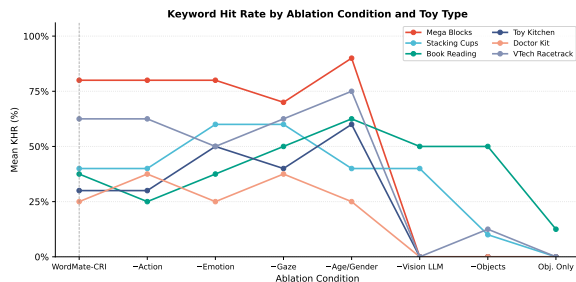


Figure 4: Keyword Hit Rate across ablation conditions for each toy type.

exact or normalised keyword matches. A contextually appropriate response using “pieces”, “goes on top”, or “make it taller” scores lower than one that mentions “blocks” by chance while describing the wrong activity entirely. KHR can also dip in the full condition when noisy automatic detections — an incorrect gaze target or a mislabelled object — are included in the prompt and the model partially follows them. Removing that noisy field can therefore produce a cleaner keyword match without improving actual scene recognition. Such occasional improvements under ablation should not be read as evidence that the removed modality is harmful; they point instead to the need for confidence weighting and conflict resolution between low-level detections and VLM interpretation.

Overall, the ablation results show that the pipeline benefits from multi-source visual context, though modalities contribute unequally. VLM enrichment provides the strongest grounding signal, and its removal causes the largest quality drop. Object detections add value, but mainly when interpreted alongside wider scene context. The uneven contribution of some modalities likely reflects the fact that the underlying perception models were not

trained for young children, caregiver–child interaction, or toy-rich settings, which contributes to noisy detections and occasional grounding errors. These findings highlight the importance of confidence weighted prompt construction and better conflict resolution between low-level detections and VLM interpretation as clear priorities for future work.

5.2 Latency Results

Figure 5 summarises latency across both platforms, Windows and Jetson, using recorded-video and live-camera inputs, covering offline processing behaviour and real-time deployment performance.

In the recorded video, the WordMate-CRI improved mean FPS from 7.2 to 29.7 on the Jetson and reduced system latency from 21.90s to 8.13s. A similar trend was observed on Windows, where FPS increased from 7.0 to 17.6 and system latency decreased from 24.14s to 10.88s. These results indicate that the cadence strategy removes unnecessary per-frame computation across both platforms. The stage-level results show that the main savings come from frame processing and vision-side computation. Other stages, particularly LLM generation and TTS, still contribute noticeably to overall delay, especially when responses are longer or speech synthesis takes more time to produce the first line. Therefore, the WordMate-CRI should be understood as a visual scheduling improvement that makes the interaction pipeline more practical on edge hardware, rather than a complete solution to all latency sources.

The live-camera provides the most important deployment evidence. On the Jetson AGX Orin, the per-frame baseline achieved only 4.3 FPS, whereas the WordMate-CRI reached 18.2 FPS, exceeding the 15 FPS live target for robot deployment. The same improvement is reflected in per-frame processing time, which dropped from 235.9 ms to 60.0 ms. This shows that the WordMate-CRI strategy changes the live system from a low-frame-rate pipeline into one that can maintain adequate temporal sampling during child play. Windows live input also improved from 2.0 to 9.5 FPS, though it did not reach the 15 FPS target, suggesting the Windows live-camera path remains more variable.

The speech pipeline results are more nuanced. On Jetson live input, speech pipeline latency was 1133.9 ms under the WordMate-CRI, compared with 1021.7 ms for the baseline. This slight increase does not undermine the cadence strategy, as the method is designed to reduce vision-side

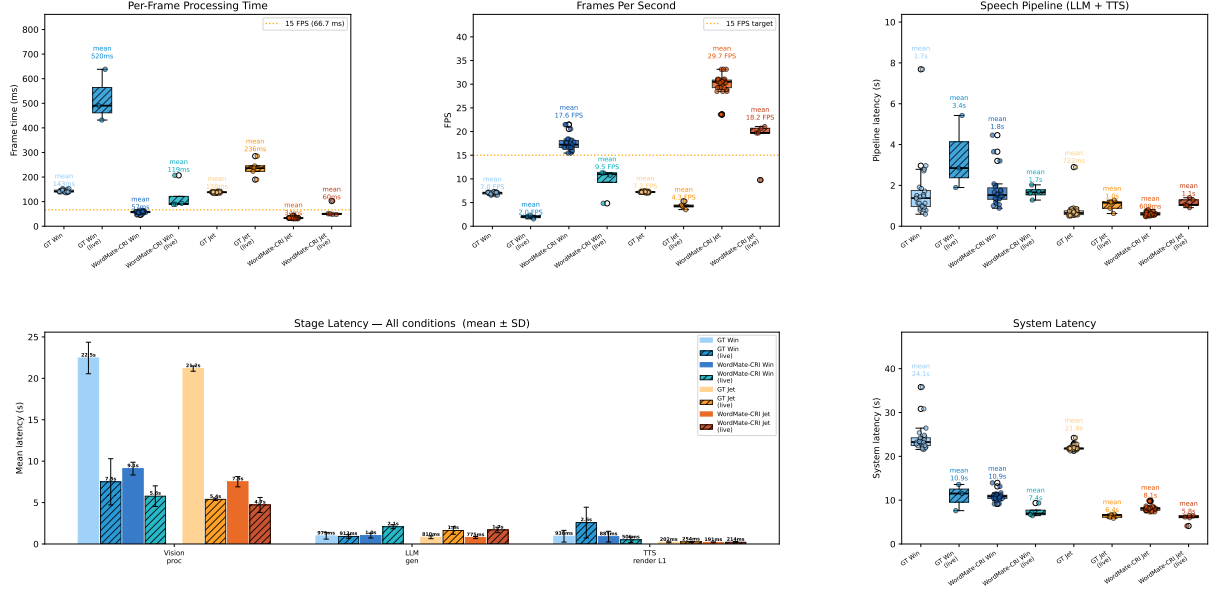


Figure 5: Latency for the per-frame baseline and WordMate-CRI on Windows and Jetson platforms. WordMate-CRI improves visual-processing throughput and enables the Jetson live deployment to exceed the 15 FPS target, while system latency is reduced despite remaining LLM and TTS costs. FPS is averaged from the reciprocal of each segment’s mean frame time. Live-condition sample counts are five records per Jetson condition and three (baseline) or four (proposed) per Windows condition; stage-latency bars show mean plus or minus one standard deviation

computation rather than shorten LLM or TTS generation. Importantly, the WordMate-CRI keeps the post-segment response delay close to one second while greatly improving live frame throughput.

On Jetson live input, system latency decreased from 6411.5 to 5844.5 ms. The smaller reduction relative to the FPS gain reflects the fixed real-time segment window: the cadence strategy does not make a five-second segment arrive earlier, but it allows the system to process it with substantially less computational pressure. The Jetson live result is therefore the key deployment finding, with the recorded-video benchmark providing supporting evidence across the six toy clips.

6 Conclusion

Building a robot that keeps up with a child’s attentional focus during play is as much a scheduling problem as it is a recognition problem. Not all information a robot needs to perceive changes at the same speed, and treating all cues as if they deserved equal computational attention is wasteful on the edge hardware that real deployments depend on. WordMate-CRI increased live Jetson throughput from 4.3 to 18.2 FPS. In the separate response evaluation, the full-context condition achieved a mean scene-grounded score of 0.967. That said, the perception models were not built with children or

toy-play in mind, and LLM generation and speech synthesis still carry latency the cadence strategy does not address. Future work should bring the perception models closer to child-centred environments and develop principled conflict resolution between low-level detections and VLM interpretation. What the results establish clearly is that temporal scheduling deserves to sit alongside model selection and multimodal fusion as a first-class design decision in child-robot perception.

6.1 Limitations

This study is an initial feasibility evaluation based on a small set of controlled toy-play scenarios. The current pipeline does not yet process children’s speech or other auditory input; augmenting the pipeline accordingly in the future will add further important information about children’s communicative intentions. Most of the perception components in the pipeline are pretrained models, including YOLO for object detection, R(2 + 1)D-18 for action recognition, OpenFace3 for gaze and facial expression analysis, 6DRepNet for head pose and MiVOLO for age and gender estimation. These models were not specifically trained for young children, caregiver–child interaction, or toy-rich play settings, which may explain some noisy detections and grounding errors. Furthermore, reference re-

sponses were reviewed by a single annotator who also inspected the system’s generated outputs during the same pass, so the evaluation is not blind to model output, and no independent inter-annotator check was performed. Future work should collect child–caregiver interaction data to refine the action model and adapt object detection to toys and materials commonly found in child-centred environments. Larger and more diverse data would also support confidence weighting, conflict resolution between low-level detections and VLM hints, and stronger evaluation in real-world CRI settings. Finally, although WordMate-CRI reduces vision-side computation, LLM generation and text-to-speech remain sources of latency that require further optimisation.

Acknowledgments

The authors wish to acknowledge generous funding by the Northwood Charitable Trust.

References

- Elkan G. Akyürek. 2025. [Temporal integration as an adaptive process in visual perception, attention, and working memory](#). *Neuroscience & Biobehavioral Reviews*, 170:106041.
- Tony Belpaeme and 1 others. 2012. [Multimodal child-robot interaction: Building social bonds](#). *Journal of Human-Robot Interaction*, 1(2):33–53.
- S. Cooper, B. Pou, A. Mayoral-Macau, A. Redondo, D. Ríos, and R. Ros. 2026. [EMOROB CARE: A low-cost social robot for supporting children with autism in therapeutic settings](#). In *Social Robotics + AI, ICSR+AI 2025*, volume 16131 of *Lecture Notes in Computer Science*, pages 607–615. Springer, Singapore.
- Timothée Dhaussy, Bassam Jabaian, and Fabrice Lefèvre. 2024. FlowAct: A proactive multimodal human-robot interaction system with continuous flow of perception and modular action sub-systems. *arXiv preprint arXiv:2408.15864*.
- J. A. Duncan, Farshid Alambeigi, and Mitch W. Pryor. 2024. [A survey of multimodal perception methods for human–robot interaction in social environments](#). *ACM Transactions on Human-Robot Interaction*, 13(4):1–50.
- Nikolaos Efthymiou, Panagiotis P. Filntisis, Petros Koutras, Anna Tsiami, James Hadfield, Gerasimos Potamianos, and Petros Maragos. 2022. [ChildBot: Multi-robot perception and interaction with children](#). *Robotics and Autonomous Systems*, 150:103975.
- Nikolaos Efthymiou, Panagiotis P. Filntisis, Gerasimos Potamianos, and Petros Maragos. 2021. [Visual robotic perception system with incremental learning for child–robot interaction scenarios](#). *Technologies*, 9(4):86.
- Diego R. Faria, A. L. Godkin, and P. P. da Silva Ayrosa. 2025. [Advancing emotionally aware child–robot interaction with biophysical data and insight-driven affective computing](#). *Sensors*, 25(4):1161.
- Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. 2008. [A hierarchy of temporal receptive windows in human cortex](#). *Journal of Neuroscience*, 28(10):2539–2550.
- D. Huang, X. Zhang, and K. Youcef-Toumi. 2026. Perceive what matters: Relevance-driven scheduling for multimodal streaming perception. *arXiv preprint arXiv:2603.13176*.
- Yuqian Jiang, Nick Walker, Minkyu Kim, Nicolas Brisonneau, Daniel S. Brown, Justin W. Hart, Scott Niekum, Luis Sentis, and Peter Stone. 2018. LAAIR: A layered architecture for autonomous interactive robots. *arXiv preprint arXiv:1811.03563*.
- Y. Kim, H. Chen, S. Alghowinem, C. Breazeal, and H. W. Park. 2023. Joint engagement classification using video augmentation techniques for multi-person HRI in the wild. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, pages 698–707, London, United Kingdom.
- Hideki Kozima, Marek P. Michalowski, and Cocoro Nakagawa. 2009. [Keepon: A playful robot for research, therapy, and entertainment](#). *International Journal of Social Robotics*, 1:3–18.
- Hideki Kozima, Cocoro Nakagawa, and Yuriko Yasuda. 2007. [Children–robot interaction: A pilot study in autism therapy](#). In *Progress in Brain Research*, volume 164, pages 385–400.
- Christopher A. Kurby and Jeffrey M. Zacks. 2008. [Segmentation in the perception and memory of events](#). *Trends in Cognitive Sciences*, 12(2):72–79.
- Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. 2011. [Topographic mapping of a hierarchy of temporal receptive windows using a narrated story](#). *Journal of Neuroscience*, 31(8):2906–2915.
- S.-R. Lu, J.-H. Lo, Y.-T. Hong, and H.-P. Huang. 2024. [Implementation of engagement detection for human–robot interaction in complex environments](#). *Sensors*, 24(11):3311.
- Ana M. G. Manea and 1 others. 2022. [Intrinsic timescales as an organizational principle of neural processing across the whole rhesus macaque brain](#). *eLife*, 11:e75540.

- Muhammad Nadeem, J. M. H. Barakat, D. Daas, and A. Potams. 2025. [A review of socially assistive robotics in supporting children with autism spectrum disorder](#). *Multimodal Technologies and Interaction*, 9(9):98.
- J. Park, P. Kim, and D. Ko. 2025. [Real-time open-vocabulary perception for mobile robots on edge devices: A systematic analysis of the accuracy-latency trade-off](#). *Frontiers in Robotics and AI*, 12.
- Brian Scassellati, Henny Admoni, and Maja Matarić. 2012. [Robots for use in autism research](#). *Annual Review of Biomedical Engineering*, 14:275–294.
- Sara E. Schroer and Chen Yu. 2023. [Looking is not enough: Multimodal attention supports the real-time learning of new words](#). *Developmental Science*, 26(2):e13290.
- Z. Shi, E. Zhao, N. Dennler, J. Wang, X. Xu, K. Shrestha, M. Fu, D. Seita, and M. Matarić. 2025. [HRIBench: Benchmarking vision-language models for real-time human perception in human-robot interaction](#). In *Proceedings of the 19th International Symposium on Experimental Robotics*.
- Alessandro Umbrico, Amedeo Cesta, Francesca Fracasso, Andrea Orlandini, and Gabriella Cortellessa. 2022. [Shaping cognitive control for HRI through the dual process theory](#). In *Proceedings of the Workshop on Cognitive Architectures for HRI*, volume 3332 of *CEUR Workshop Proceedings*.
- Alessandro Umbrico, Riccardo De Benedictis, Francesca Fracasso, Amedeo Cesta, Andrea Orlandini, and Gabriella Cortellessa. 2023. [A mind-inspired architecture for adaptive HRI](#). *International Journal of Social Robotics*, 15:371–391.
- Yuichi Yamashita and Jun Tani. 2008. [Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment](#). *PLoS Computational Biology*, 4(11):e1000220.
- Chen Yu, Sumarga H. Suanda, and Linda B. Smith. 2019. [Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months](#). *Developmental Science*, 22(1):e12735.
- W. Zhao, K. Gangaraju, and F. Yuan. 2025. [Multimodal perception-driven decision-making for human-robot interaction: A survey](#). *Frontiers in Robotics and AI*, 12:1604472.