

Seeing Only When Needed: A Multimodal Architecture for Context-Aware Social Dialogue

Maria Jose Pinto Bernal
IDLab AIRO, Ghent University-imec
Ghent, Belgium
mariajose.pintobernal@ugent.be

Tony Belpaeme
IDLab AIRO, Ghent University-imec
Ghent, Belgium
tony.belpaeme@ugent.be

Abstract

Large Vision–Language Models are increasingly used for visually grounded social dialogue, yet most systems assume that vision should be active continuously, adding computational load and increasing the risk of unnecessary or hallucinated descriptions. We present a multimodal architecture that treats vision as a selective, context-dependent resource. A lightweight vision-gating module triggers visual grounding only when a user utterance requires it, while a complementary ambient monitoring component detects gradual scene changes at a low frame rate. Both pathways contribute cues only when relevant, enabling the robot to use visual information meaningfully without overuse. A preliminary evaluation with 10 participants (≈ 95 minutes) shows that the gating mechanism identified vision-relevant turns with 93.4% accuracy, and that grounded descriptions aligned with the scene in 90.7% of cases.

CCS Concepts

• **Human-centered computing** → **Interaction design process and methods.**

Keywords

multimodal dialogue, social robots, vision-language models, context-aware interaction, human–robot interaction.

ACM Reference Format:

Maria Jose Pinto Bernal and Tony Belpaeme. 2026. Seeing Only When Needed: A Multimodal Architecture for Context-Aware Social Dialogue. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26)*, March 16–19, 2026, Edinburgh, Scotland, UK. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3776734.3794486>

1 Introduction

In recent years, integrating Vision-Language Models (VLMs) has opened new possibilities for human–robot interaction (HRI) due to their ability to generate contextually relevant textual descriptions from visual inputs [11]. These models are characterised by their effectiveness in capturing and translating complex visual patterns into coherent linguistic representation [5]. This capability

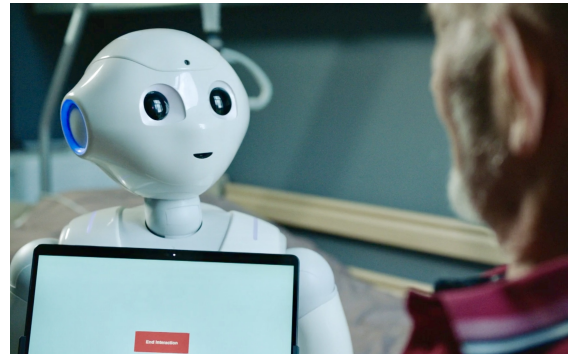


Figure 1: **Experimental setup.** Pepper interacting with a participant; an external Microsoft Surface tablet is mounted on the robot.

promises more flexible and context-aware interaction than traditional scripted dialogue systems, which relied on separate pipelines for vision and speech.

However, enabling a robot to “see” during interaction also introduces a new design question. In most current systems, perception is always active, i.e., every user message is paired with an image or video frame, and visual context is passed automatically to the language model. While effective in controlled demonstrations, this assumption becomes problematic in everyday conversation. Not every exchange requires visual grounding, and using vision indiscriminately can cause the robot to introduce unnecessary descriptions or shift attention away from the user’s intended topic [9, 10]. Moreover, constantly processing images adds computational overhead and constrains responsiveness, especially in real-time deployments.

This observation raises a simple but overlooked question: *Does a social robot need to see all the time?* A natural point of comparison is how humans manage visual attention during conversation. Human visual attention is inherently selective and context-dependent rather than continuous. Conversation studies show that people rely on vision at specific, meaningful moments—such as checking a partner’s reaction or coordinating turn transitions, but often look away when concentrating, formulating speech, or when visual input is unnecessary [14, 17]. This pattern reflects a “just-in-time” strategy in natural vision, where humans gather visual information precisely when it is relevant to the current task, not continuously [2, 7]. Applying this principle to social robots may support more natural interaction while reducing unnecessary visual computation.

In this work, we introduce a context-aware vision-gating mechanism that enables a robot to decide when visual information is relevant to the ongoing conversation. Instead of supplying vision on every turn, the system analyses each user message and activates



This work is licensed under a Creative Commons Attribution 4.0 International License. *HRI Companion '26, Edinburgh, Scotland, UK*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2321-6/2026/03
<https://doi.org/10.1145/3776734.3794486>

visual reasoning only when the conversational context requires it. When vision is not relevant, the robot proceeds in text-only mode, avoiding unnecessary image processing and preventing irrelevant visual references.

Our architectural design was motivated by observations made during real-world multimodal deployments, where providing visual input at every turn frequently added processing overhead, reduced system stability, and introduced irrelevant visual details that acted as "noise" in the dialogue, shifting the robot away from the user's intended topic and disrupting the conversation flow. Treating visual perception as a selective, context-dependent component allowed us to avoid these issues while maintaining access to visual grounding when it was genuinely needed. Our contribution is not a new model, but an architectural approach that prioritises controlled and context-aware visual integration to support stable and efficient multimodal dialogue in real-world HRI settings.

2 Related Work

The integration of perception and dialogue has long been a central goal in HRI. Early social robots typically combined separate modules for object recognition, speech processing, and dialogue management, resulting in limited interaction between what the robot perceived and what it said [3]. Recent advances in VLMs, however, have begun to dissolve this separation by enabling unified reasoning over multimodal inputs, thereby supporting open-ended, visually grounded dialogue.

Recent studies have begun integrating VLMs directly into interactive social robots. Early approaches combined image-to-text models with an LLM to generate visually aware conversation [8], but multimodal foundation models have been shown to produce visually grounded conversational responses by adding images directly to the conversation prompt [1]. Others deploy lightweight versions, such as LLaVA-OneVision, to generate context-aware greeting utterances from short video sequences in real time [4]. Personalised approaches, such as USER-VLM 360°, incorporate user-specific visual embeddings to tailor social exchanges [13]. Additional research combines VLM-based scene understanding with LLM reasoning to infer whether initiating interaction is socially appropriate [15]. Together, these works highlight a growing interest in unified perception-dialogue pipeline.

Despite this progress, integrating visual processing directly into the conversational loop introduces several challenges. VLMs often over-describe a scene, producing outputs that resemble captions rather than socially calibrated responses. Incorporating visual tokens also increases computational load, which can degrade responsiveness in real-time interaction. In addition, VLMs often hallucinate objects or attributes [10, 18], with grounding errors persisting even in leading models [11, 16]. These issues are particularly problematic in HRI, where delays or mis-grounded utterances can disrupt turn-taking and reduce user trust.

What remains largely underexplored is when a robot should engage its visual perception. Most existing systems assume continuous visual availability and rarely model whether vision is appropriate or necessary for a given conversational turn. This creates a gap between technical capability and interactional appropriateness. To address this, we propose a lightweight vision-gating mechanism

that determines when visual reasoning is relevant, treating perception as a selective resource rather than a constant stream. By conditioning vision on conversational need, the system aims to reduce unnecessary visual output and computational overhead while supporting more contextually appropriate multimodal dialogue.

3 System Overview

We developed a multimodal conversational architecture, that was implemented on the Pepper robot, designed to integrate language, vision, and dialogue management within a single application running on an external device. The setup replaces Pepper's internal tablet with a Microsoft Surface, mounted using a custom 3D-printed stand (Figure 1). This configuration was chosen because Pepper's native software stack limits integration of modern multimodal pipelines. Running the system on the Surface allows us to use current LLM and VLM models and maintain control over timing, privacy, and module design, while Pepper continues to provide embodiment through gestures, face tracking, head movements, LEDs, and audio playback.

The overall architecture operates as a sequence of tightly coordinated modules (Figure 2). The system continuously processes audio through a turn-taking component that monitors speech activity, silence, and semantic cues to determine when the user's turn has ended, this was based on a previous study reported in [12]. Once the user's turn is detected, the final ASR transcription is passed to the context-integration stage, in which it evaluates whether additional perceptual or contextual information is needed. This is achieved through independent decision modules that contribute information only when relevant.

A key feature of the system is its dual visual-context architecture, consisting of (i) an on-demand vision-gating mechanism for grounding individual user turns, and (ii) a lightweight ambient visual monitoring module that provides subtle contextual cues when relevant. These modules serve complementary functions and operate at different temporal and computational scales but both serve to ensure that vision is used selectively rather than continuously.

The **task-oriented vision-gating module** is triggered only when the system suspects that a user's utterance requires visual grounding. To avoid latency from real-time image capture, the Surface continuously maintains a rolling buffer of the most recent camera frames (5 last frames). If the gating agent determines that vision is needed, the system retrieves these buffered frames and sends them to a VLM (GPT-4o) together with the user's utterance. The resulting grounding summary is then incorporated into the dialogue context. If grounding is unnecessary, the VLM is not called, reducing computational cost and avoiding irrelevant visual references that could introduce noise into the dialogue.

The **ambient visual monitoring module**, in contrast, runs asynchronously at a low frame rate and is intentionally designed to be lightweight. Frames are sampled at a low rate of 0.5 fps, which is sufficient for detecting gradual ambient changes while avoiding interference with the robot's primary speech and dialogue processes. Instead of relying on a large cloud-based VLM, this module uses a local model—in our implementation, a compact VLM such as MiniGPT-4 [6] or an equally efficient vision encoder. We selected this type of model not only for its modest computational

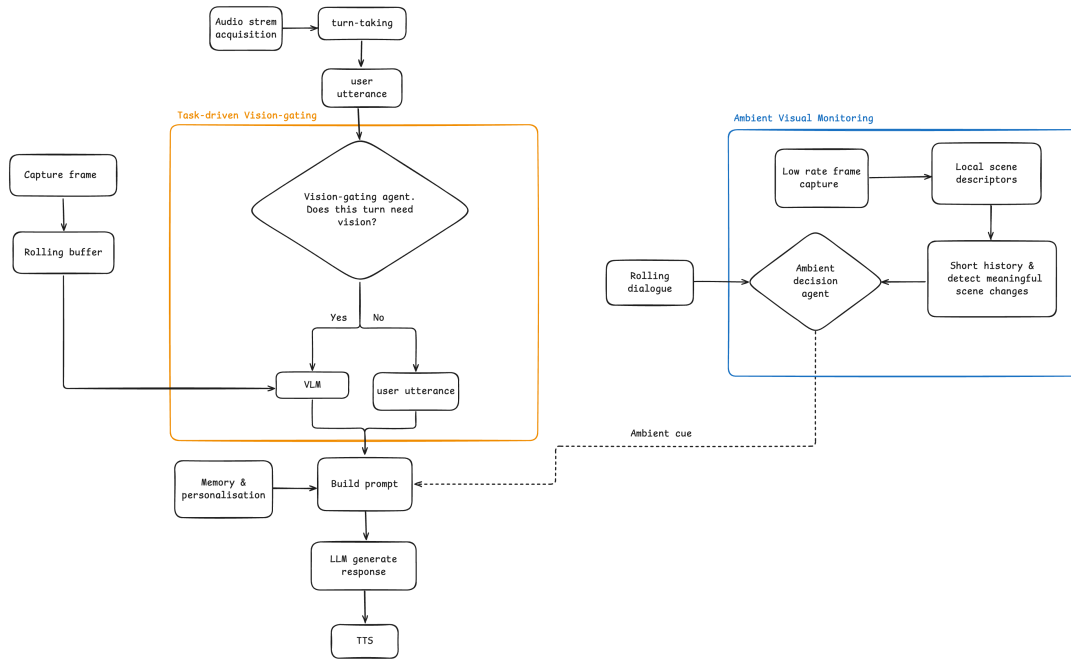


Figure 2: **Proposed multimodal dialogue architecture.** User speech is processed by a turn-taking module and passed to a vision-gating agent that determines whether visual grounding is required. When needed, the system retrieves the most recent frames from a rolling buffer and invokes a VLM to generate a concise grounding summary; otherwise, the dialogue proceeds in text-only mode. In parallel, a lightweight ambient visual monitoring module samples the scene at a low frame rate to detect gradual environmental changes. Ambient cues are only forwarded when sustained changes are detected and deemed contextually relevant. Selected contextual information is assembled by the prompt builder before generating a spoken response.

demands, but also because recent evaluations show that MiniGPT-4 exhibits substantially fewer object hallucinations than many other LVLMS [18]. This reduced tendency to introduce incorrect visual details is especially important for ambient monitoring, where false detections could unnecessarily disrupt the dialogue.

For each sampled frame, the ambient module generates a short caption. Rather than storing all captions, the system maintains a small rolling buffer and compares each new caption to the previous one using sentence-level semantic similarity. If the similarity is high (≥ 0.90), the previous caption is simply replaced, preventing redundant entries. When similarity drops below a threshold (≤ 0.80), the system flags a potential scene change and performs a lightweight comparison between captions (e.g., new or disappearing objects, arrival of another person, posture changes). Only if this difference persists across frames, indicating that it reflects an actual scene change rather than VLM fluctuation, is it promoted to an event-level cue and passed to the ambient decision agent, which determines whether the cue is contextually relevant and helpful for the ongoing interaction.

Once the task-driven and ambient decision mechanisms have determined which perceptual cues are relevant, the selected information is passed directly to the prompt builder. The prompt builder does not perform additional filtering; it simply assembles the contextual elements that have already been selected—including grounding summaries from the task-driven vision module, ambient cues when

they arise, the user’s utterance, the rolling conversational history, and the user’s personalisation and memory data. Importantly, the dialogue model receives visual information only after it has been evaluated by these upstream mechanisms; it does not decide when vision should be triggered or which perceptual cues matter. This design keeps visual processing and language generation separate, ensuring that contextual signals are incorporated in a controlled and predictable way.

To produce coherent and personalised responses, the prompt builder composes a structured instruction for the LLM. This prompt integrates: (i) the robot’s persistent persona, (ii) user-specific information such as profile, language, and memory summaries, (iii) the user’s latest utterance and the conversational context, (iv) visually grounded information from VLM when required, and (v) ambient cues from the local monitoring module when contextually relevant. The persona block establishes warmth and continuity, while the behavioural guidelines promote natural, varied, and empathetic dialogue. For example, the guidelines instruct the model to adapt its tone to the user’s emotional state, reference past topics naturally without looping, and introduce gentle follow-ups that connect meaningfully to the user’s previous statements. Once assembled, the prompt is submitted to the LLM (GPT-4o), which generates the robot’s response and then the reply is then synthesised through ElevenLabs Text-to-Speech and delivered to Pepper.

4 Preliminary Evaluation

To examine whether the proposed architecture behaves as intended during real interaction, we conducted a preliminary evaluation focusing on technical performance and the qualitative behaviour of the system’s key modules. Our goal was to assess whether the system could sustain fluid, grounded dialogue while triggering visual reasoning only when contextually appropriate, thereby preserving coherence and avoiding unnecessary or speculative visual interpretations.

We conducted a series of short pilot conversations with 10 adult participants (6 female, 4 male; $M_{\text{age}} = 28.7$, $SD = 3.88$) each of whom engaged in a 10 – 12 minutes free conversation with the robot. Participants were invited to speak naturally, move freely, and interact with objects present in the environment (e.g., mugs, photo albums, books, or personal items) without being given specific tasks. The sessions took place in a quiet open room that remained accessible to researchers, meaning that occasional background events—such as someone briefly entering the space or walking past the doorway—occurred naturally but without causing disruption. These mild variations provided a realistic setting for evaluating both the task-oriented gating module and the ambient change-detection mechanism, while maintaining a controlled, conversational environment overall.

Across all sessions, the interactions resulting in approximately 95 minutes of dialogue, yielding a total of 537 user turns. System logs captured turn boundaries, gating decisions, VLM calls, ambient descriptors, and timing information, while all video recordings were later annotated independently by two coders to assess the correctness and appropriateness of the system’s visual reasoning.

A central question for this preliminary evaluation was whether the vision-gating agent correctly identified when visual grounding was genuinely required. Annotators judged that 82 of the 537 user turns (15.27%) contained explicit or implicit references to objects or visually observable situations (e.g., “Do you like this mug?”, “What am I doing here?”). The gating agent activated in 76 of these cases, producing a gating accuracy of 93.4%. When activated, the system retrieved the last five frames from the rolling buffer and sent them, together with the user’s utterance, to the VLM to obtain a concise visual interpretation. Annotators compared these interpretations with the recorded scenes and found that 69 of the 76 grounded turns (90.7%) accurately captured the relevant visual elements. The remaining interpretations showed mild overspecification or generic phrasing, such as describing a plain mug as “a ceramic mug with a slightly glossy finish” or referring to a photo album as “a book-like object.” These instances were still consistent with the visible scene and did not introduce inaccuracies or unwanted topic shifts. In contrast, many grounded turns demonstrated precise visual alignment, such as correctly identifying a “red ceramic mug,” recognising a “photo album on the participant’s lap,” or detecting that “another person entered the frame from the right.”

A small number of activations ($\approx 12.86\%$), however, were triggered by linguistic patterns that the gating agent interpreted as potentially requiring visual grounding even though the referent was ambiguous or ultimately irrelevant. These language-induced activations, typically involved utterances containing deictic expressions such as “this,” “here,” or “what about this?” without a

sufficiently clear object being specified in the dialogue context. When the gate activated in these cases, the VLM returned a neutral, scene-consistent interpretation of the most recent frames, and the main LLM correctly treated this information as non-essential. Importantly, the robot did not introduce visual details that were not present, fabricate objects, or shift the conversation toward unintended topics indicating, that even overly cautious gating decisions remained safe and non-disruptive.

The complementary pattern occurred when participants gestured toward something in the room but produced incomplete or vague utterances that lacked sufficient linguistic grounding (e.g., “What’s that?” without establishing a referential anchor). In these situations, the gating mechanism remained inactive, and the robot responded based on dialogue context alone.

Regarding the ambient visual monitoring module, the system behaved conservatively but consistently across all sessions. Across the 95 minutes of interaction, the module detected 35 sustained changes in the environment, such as participants placing or removing objects on the table, adjusting their posture, momentarily leaving the frame, or another person briefly entering the scene. These detections are based entirely on coarse local features (e.g., changes in object silhouettes, new motion clusters, head/torso orientation shifts) rather than detailed semantic scene interpretation. Of the 35 detected changes, the ambient decision agent surfaced 14 cues as contextually relevant. These cues were typically associated with moments when the participant’s behaviour shifted in a way that could support conversational continuity. For example, when a participant began browsing a photo album during a lull in dialogue, or when another person briefly entered the room. In these cases, the robot incorporated the cue naturally (e.g., “I see you’re looking at something. Are those photos special to you?”), which participants described as appropriate and timely.

Importantly, transient fluctuations—such as momentary hand movements or brief posture adjustments—were filtered out and never passed to the dialogue model. When no meaningful change occurred, the ambient module correctly refrained from contributing any information, preventing unnecessary background descriptions or noise in the prompt. Across all sessions, ambient cues remained infrequent, subtle, and non-disruptive, aligning with the intended design of maintaining low-intrusion environmental awareness.

5 Conclusion

We presented a multimodal dialogue architecture that treats vision as a selective, context-dependent resource by combining on-demand visual grounding with lightweight ambient monitoring. A preliminary validation showed that the system can reliably identify when visual information is required, while avoiding unnecessary perception, hallucinated descriptions, and disruptive shifts in dialogue.

The current evaluation primarily verifies architectural feasibility and functional behaviour, rather than users’ subjective perceptions or comparative benefits over continuous visual input. Gesture-based referents without linguistic anchors also remain challenging, reflecting the limits of language-driven gating. Future work will therefore examine interactional outcomes, user trust, and error recovery, as well as direct comparisons between selective and continuous perception strategies in longer-term deployments.

References

- [1] Giulio Antonio Abbo and Tony Belpaeme. 2025. I was blind but now i see: Implementing vision-enabled dialogue in social robots. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1176–1180.
- [2] Dana H Ballard, Mary M Hayhoe, and Jeff B Pelz. 1995. Memory representations in natural tasks. *Journal of cognitive neuroscience* 7, 1 (1995), 66–80.
- [3] Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67–67.
- [4] Giuseppe De Simone, Loris Roveda, Alessia Saggese, and Mario Vento. 2025. Towards Proactive Social Robots: Distilling Visual Knowledge from Large Vision-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2492–2502.
- [5] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).
- [6] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790* (2023).
- [7] Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9, 4 (2005), 188–194.
- [8] Ruben Janssens, Pieter Wolfert, Thomas Demeester, and Tony Belpaeme. 2024. Integrating visual context into language models for situated social conversation starters. *IEEE Transactions on Affective Computing* 16, 1 (2024), 223–236.
- [9] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13872–13882.
- [10] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 292–305. doi:10.18653/v1/2023.emnlp-main.20
- [11] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A Survey of State of the Art Large Vision Language Models: Benchmark Evaluations and Challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1587–1606.
- [12] Maria J Pinto and Tony Belpaeme. 2024. Predictive turn-taking: Leveraging language models to anticipate turn transitions in human-robot dialogue. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 1733–1738.
- [13] Hamed Rahimi, Adil Bahaj, Mouad Abrini, Mahdi Khoramshahi, Mounir Ghogho, and Mohamed Chetouani. 2025. User-vm 360: Personalized vision language models with user-aware tuning for social human-robot interactions. In *Proceedings of the 27th International Conference on Multimodal Interaction*. 326–336.
- [14] Federico Rossano. 2012. Gaze in conversation. *The handbook of conversation analysis* (2012), 308–329.
- [15] Kazuhiro Sasabuchi, Naoki Wake, Atsushi Kanehira, Jun Takamatsu, and Katsushi Ikeuchi. 2025. Agreeing to interact in human-robot interaction using large language models and vision language models. *arXiv preprint arXiv:2503.15491* (2025).
- [16] Ashish Seth, Dinesh Manocha, and Chirag Agarwal. 2025. HALLUCINOGEN: Benchmarking Hallucination in Implicit Reasoning within Large Vision Language Models. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*. 89–102.
- [17] Jack Sidnell. 2006. Coordinating gesture, talk, and gaze in reenactments. *Research on Language and Social Interaction* 39, 4 (2006), 377–409.
- [18] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754* (2023).

Received 2025-12-08; accepted 2026-01-12; revised 8 December 2025; revised 4 January 2026; accepted 12 January 2026