Semantic Glitch: Agency and Artistry in an Autonomous Pixel Cloud

Qing Zhang
The University of Tokyo
qzkiyoshi@gmail.com,

Jing Huang
Tokyo University of the Arts
hkoukenj@gmail.com,

Mingyang Xu
Keio University
mingyang@kmd.keio.ac.jp,

Jun Rekimoto
The University of Tokyo
SONY CSL Kyoto
rekimoto@acm.org

Abstract

While mainstream robotics pursues metric precision and flawless performance, this paper explores the creative potential of a deliberately "lo-fi" approach. We present the "Semantic Glitch," a soft flying robotic art installation whose physical form, a 3D pixel-style cloud, is a "physical glitch" derived from digital archaeology. We detail a novel autonomous pipeline that rejects conventional sensors like LiDAR and SLAM, relying solely on the qualitative, semantic understanding of a Multimodal Large Language Model to navigate. By authoring a bio-inspired personality for the robot through a natural language prompt, we create a "narrative mind" that complements the "weak," historically loaded body. Our analysis begins with a 13-minute autonomous flight log, and a follow-up study statistically validates the framework's robustness for authoring quantifiably distinct personas. The combined analysis reveals emergent behaviors, from landmark-based navigation to a compelling "plan-to-execution" gap, and a character whose unpredictable, plausible behavior stems from a lack of precise proprioception. This demonstrates a lo-fi framework for creating imperfect companions whose success is measured in character over efficiency.

1 Introduction

In an era where digital imagery relentlessly pursues high fidelity, why has the "pixel" aesthetic, born from technical limitations, sparked a persistent wave of retro-futurism? [5, 18] Furthermore, when a symbol composed of pixels, which should exist on a two-dimensional screen, suddenly acquires a physical body and floats among us like a seemingly autonomous creature, how does our relationship with it, and our perception of the virtual and the real, change? [22, 17] This paper explores these questions by detailing the creation and behavior of the "Pixel Cloud," a soft robotic art installation that gains its physical autonomy from a Multimodal Large Language Model (MLLM). Our approach to authoring an agent's character builds on artistic and scientific explorations into crafting lifelike [3, 4], emergent behaviors for interactive robotic agents. This work does not aim to solve any practical problem. Instead, it follows the "Speculative Design" philosophy advocated by Anthony Dunne and Fiona Raby [2], functioning as a "speculative object" to provoke public imagination and debate. It poses a series of "what if" questions: What if the untouchable digital "cloud" had a visible, fragile, physical body? What if the symbols from our digital childhood memories gained physical autonomy? By combining media archaeology [18] with robotics, the core thesis of this work is that through a "physical hack" [21] of the pixel, we can reveal and reshape the increasingly complex "entangled agencies" [23] among humans, machines, and the environment.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

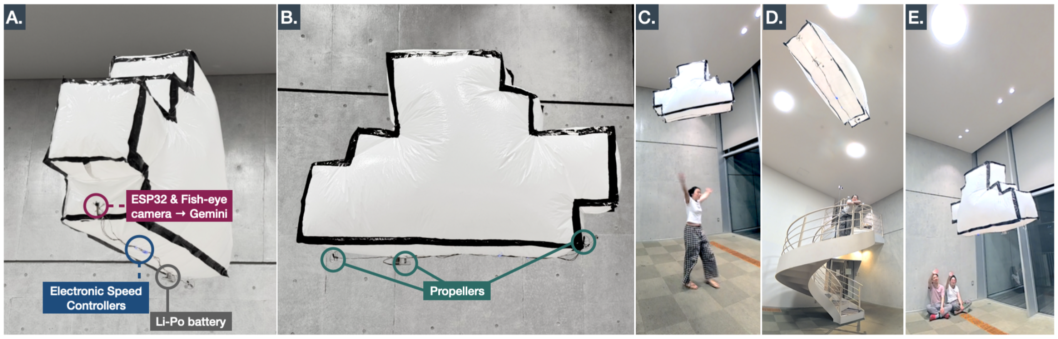

Figure 1: The "Semantic Glitch" hardware, flight behavior, and first-person perspective with MLLM-generated reasoning. (A, B) The robot's body, showing the placement of the ESP32 with a fish-eye camera, electronic speed controllers, Li-Po battery, and propeller modules. (C, D, E) The robot in flight, demonstrating its interaction with the environment and its "perspective-dependent morphological illusion."

Grounded in media archaeology and speculative design [20, 19], this paper traces the "Pixel Cloud" from its "physical glitch" body to its narrative AI mind, then analyzes both a detailed autonomous flight case study and an expanded persona validation. We conclude by considering how this "lo-fi" approach reframes robotic agency around character, plausibility, and relational presence.

2 The Body: A Deliberate "Physical Glitch"

The robot's physical form is a deliberate "physical glitch," designed to embody the "Yowai Robotto" (Weak Robot) philosophy by rejecting metric precision in favor of character [7, 8, 10, 12]. A core engineered feature is its "perspective-dependent morphological illusion": from one angle, it appears as a 2D pixel image, but as it rotates, its 3D voxel structure is revealed (Fig. 1 C-E). This effect translates a software "error" into a tangible, repeatable imperfection. Constructed as a soft, fragile helium blimp, its form is intentionally "weak" to invite empathetic interaction [12, 13]. As later sections show, this physical weakness also sets up a productive tension between semantic intention and embodied execution.

3 The Mind: Navigation as Bio-Inspired Narrative

Rejecting Metric Precision: The conventional path to robotic autonomy involves building a precise, mathematical model of the world. This is typically achieved with a suite of metric sensors (such as LiDAR or infrared depth sensors) and complex algorithms like SLAM (Simultaneous Localization and Mapping), a technology envisioned as a future step in the project's initial conceptualization. We deliberately rejected this path. A SLAM-based robot, with its metric geometric understanding, would be philosophically "out of character." Its calculated, optimal movements would be incongruous with the artifact's ephemeral nature, breaking the illusion of an animate entity. Therefore, to maintain the "weak robot" concept, the mind's perception had to be as abstract as the physical form.

The "Lo-Fi" Semantic Engine: In place of a complex, sensor-heavy system, we embraced a framework of stateful semantic reasoning. The agent's autonomy is powered by a novel, two-stage "lo-fi" pipeline that separates global scene understanding from local decision-making.

The entire control loop is orchestrated by a host computer (MacBook Pro, M4 Max, 64GB RAM) running a Python script, which communicates with the robot's XIAO ESP32S3 core. The ESP32S3 is responsible only for low-level tasks: streaming video from its camera (160° fisheye lens) and actuating its propellers via WebSocket commands. All cognition is offloaded to the remote Gemini 2.5 Flash API, subject to its terms of use, transforming it into a stateful, MLLM "mind."

This is achieved through two distinct phases, as illustrated in Figure 2. First, the **Preamble Stage** performs zero-shot spatial mapping. Upon initialization, the system begins a stateful ChatSession with the Gemini API. It sends the PREAMBLE_PROMPT along with a single 360° panoramic image

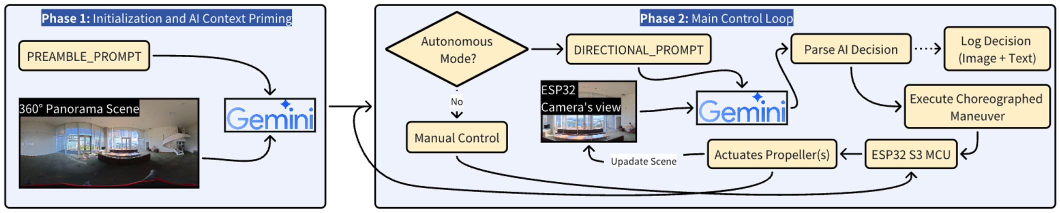

Figure 2: The two-phase semantic reasoning pipeline. Phase 1 (Initialization): A single 360° panorama and a PREAMBLE_PROMPT are sent to the Gemini API to establish a stateful "mental map". Phase 2 (Control Loop): In a continuous loop, the system uses the live camera view and a DIRECTIONAL_PROMPT to generate context-aware actions, which are then logged and executed by the robot.

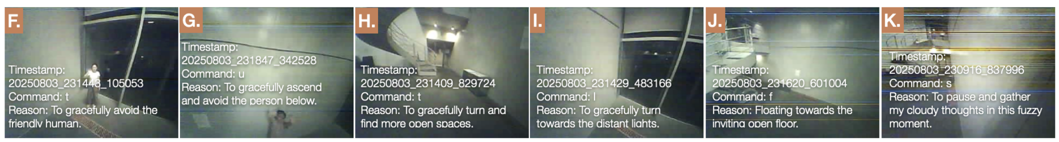

Figure 3: Key moments from the agent's first-person perspective, linking its visual input to its logged decisions. (F) Lateral avoidance of a person. (G) Vertical avoidance of a person. (H) A corrective turn near the staircase, illustrating the plan-to-execution gap. (I) Goal-oriented navigation towards distant lights. (J) Seeking open space. (K) A moment of contemplative inaction.

of the environment. This one-time action tasks the AI with performing a high-level analysis of the entire operational area, identifying boundaries, major landmarks, open fly-zones, and obstacles before any physical movement occurs. This establishes a persistent "mental map" within the AI's chat context, a process that took 2.81 seconds to complete during our experiment.

Second, the **Directional Stage** handles context-aware deliberation. In the main operational loop, we recast navigation as a continuous Visual Question Answering (VQA) problem [11]. For each decision, the DIRECTIONAL_PROMPT is posed as the "question," which the MLLM "answers" by interpreting the "vision" context from the ESP32S3's live video frame. The resulting answer, containing both a command and a narrative reason, is informed by the global spatial map established in the preamble stage. This creates a continuous, state-aware feedback loop where local perception is fused with global memory. During operation, this decision loop exhibited a mean latency of 2.8 ± 0.3 seconds, quantitatively defining the agent's deliberate, non-continuous cognitive cycle.

Prompt Engineering for Hierarchical Cognition: The agent's hierarchical cognitive process is authored entirely through two carefully engineered natural language prompts. These prompts define the function of each stage of the reasoning pipeline.

The PREAMBLE_PROMPT serves as the high-level cartographer, instructing the model to deconstruct the panoramic scene into a structured, semantic map. The DIRECTIONAL_PROMPT acts as the low-level navigator. It instructs the agent to use its established "mental map" as prior knowledge when interpreting the live camera feed to make an immediate choice. The prompt explicitly defines the agent's available actions; the full, verbatim text for both the PREAMBLE_PROMPT and DIRECTIONAL_PROMPT is provided in Appendix A.

This two-prompt structure fundamentally changes the nature of navigation. The AI does not simply react to pixels or hard-coded logic; it situates its local perception within a persistent, global narrative, allowing for more sophisticated, long-term reasoning without the overhead of conventional mapping algorithms. Ultimately, this pipeline serves as a new model for authoring complex AI behavior, where high-level strategic goals and low-level personality traits can be defined and layered through natural language [16]. By representing all outputs as text, our work aligns with a broader trend of using language as a unified interface for robotic control, though we apply it here to generate character-rich narrative instead of precise coordinates.

4 Analysis of Emergent Dialogue: A Case Study in Stateful Navigation

This section analyzes a 13-minute continuous operational log to demonstrate how the synthesis of a lo-fi form and a stateful mind produces a uniquely plausible form of artificial life, using examples illustrated in Figure 3.

Goal-Oriented Navigation and Landmark Use: The logs confirm that the Preamble stage was successful. The agent consistently uses landmarks identified in the 360° panorama to inform its navigation, demonstrating effective use of its "mental map." This is evident in decisions where it actively seeks out desirable zones, such as when turning with the reason `''t, To gracefully turn towards the distant lights''` (Fig. 3-I) or moving forward because it is `''f, Floating towards the inviting open floor''` (Fig. 3-J). The spiral staircase was a particularly prominent landmark, with the agent often reasoning about it as a point of interest (e.g., `''l, To drift away from the wall and admire the elegant spiral''`). This behavior shows the agent is not just reactively avoiding obstacles but is executing long-term, goal-oriented exploration based on the initial context.

Emergent Social Robotics and Character: The "bio-inspired" personality authored in the prompts resulted in sophisticated social behaviors. The agent recognized and reacted to humans in a way that was consistent with its prompt-defined nature. All interactions involving human observers were conducted under a protocol approved by our university's Institutional Review Board. **Dynamic Human Avoidance**: The agent employed varied strategies when encountering people. On one occasion, it chose a lateral maneuver: `''t, To gracefully avoid the friendly human''` (Fig. 3-F). In a different situation, it opted for a vertical solution: `''u, To gracefully ascend and avoid the person below''` (Fig. 3-G). This demonstrates a flexible decision-making capability, choosing different actions for similar problems based on the specific context. **Contemplative Behavior**: The agent's character was further revealed in moments of inaction. The log contains entries like `''s, To pause and gather my cloudy thoughts in this fuzzy moment''` (Fig. 3-K). These are not error states; they are authored behaviors that give the agent a plausible, non-utilitarian, and creature-like quality, directly supporting the "Yowai Robotto" concept.

The Plan-to-Execution Gap and the "Yowai Robotto": The most notable "glitchy" behaviors arise from the conflict between the agent's high-level semantic understanding and its lack of low-level physical self-awareness. The agent knows what it wants to do but not precisely how to do it. For instance, after successfully navigating near the staircase, it decides to make a corrective turn with the command `''t, To gracefully turn and find more open spaces''` (Fig. 3-H). The image shows it is close to the structure, and while the intention is correct, the subsequent maneuver is clumsy. It lacks the precise proprioceptive knowledge of its own momentum or the physical dynamics of its forward-arcing maneuver. This constant "wrestling" with its own physical form—a direct result of the plan-to-execution gap—is a clear and authentic expression of a "weak" agent that feels organic rather than programmed.

The Voice of the Glitch: Narrative as an Artistic Medium: Beyond the dialogue expressed through physical movement, the "Semantic Glitch" communicates through a third, crucial modality: its own generated text. The simple string of text that accompanies each action—such as `''l, To drift away from the wall and admire the elegant spiral''` or `''s, To ponder the shimmering, uncertain view''`—functions as more than a mere debug log or explanation. It is a performance of an "internal monologue," a textual broadcast from the agent's narrative core. This output can be contrasted with the functional "Chain-of-Thought" (CoT) [15] reasoning used in performance-oriented driving models to generate an explicit driving rationale for improved safety and accuracy. In our work, this textual output is not a means to a more accurate end, but a distinct artistic medium in itself, what can be interpreted as a form of minimalist, AI-driven poetry that completes the artifact's character.

The poetics of this voice are "lo-fi" by design, mirroring the aesthetics of the body and mind. The phrases are short, direct, and laden with qualitative, emotional language. Rather than using simple objective descriptions, the agent's vocabulary is rich with evocative verbs like "admire," "embrace," "pirouette," and "ponder." Similarly, the moments of contemplative hesitation, such as `''s, To pause and gather my cloudy thoughts...''`, are a potent performance of the vulnerability central to the "Yowai Robotto" concept. It is through this textual voice that the agent's designed personality—its shyness, curiosity, and uncertainty—is communicated most explicitly to the audience.

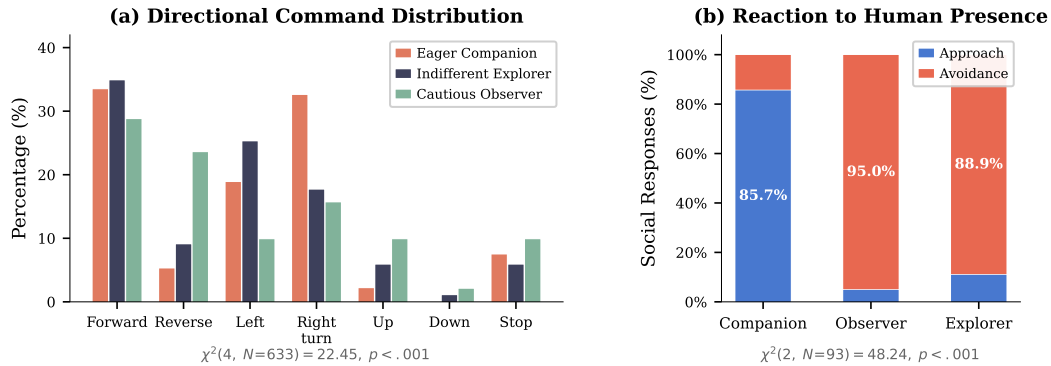

Figure 4: Behavioral analysis across personas. **(a)** Distribution of directional commands issued by Gemini across three personas ($\chi^2(4,\ N{=}633) = 22.45,\ p < .001$). **(b)** Social stance when humans are detected: approach vs. avoidance responses ($\chi^2(2,\ N{=}93) = 48.24,\ p < .001$).

This completes the artistic triad of the work. The full experience of the "Semantic Glitch" is produced by the synthesis of: The Form: The unstable, pixelated body that acts as a "physical glitch." The Movement: The pulsating, character-rich wandering that results from a lack of proprioception. The Voice: The minimalist, poetic text that reveals the agent's internal, narrative state.

Ultimately, the dialogue is not just between the cloud's mind and its body, but between the complete agent and the viewer. The agent's voice gives the audience a direct window into its unique perception of their shared space, blurring the boundary between the agent's world and their own. It is this final layer of communication that transforms the artifact from a fascinating machine into a coherent, multi-layered character with which an audience is invited to empathize.

5 Discussion

The case study results from our analysis demonstrate that a "lo-fi" symbiotic agent can achieve sophisticated and character-rich autonomous behavior. This section discusses the implications of these findings, focusing on our method as a model for creative AI, the role of limitations in generating character, and the broader impact on the field of human-robot interaction.

Validation of Robustness and Authorial Control: To address the limitations of a single case study and the critique of "light validation," we conducted an expanded study to test the framework's robustness across multiple authored personas and environments. We authored three distinct personalities—an "Eager Companion" (pro-social), a "Cautious Observer" (avoidant), and an "Indifferent Explorer" (neutral)—and tested them in two different indoor locations. As shown in Figure 4a, the authored prompts produced distinct and statistically significant "behavioral fingerprints" in the distribution of directional commands issued by Gemini ($\chi^2(4, N = 633) = 22.45, p < .001$), demonstrating that the personas were not merely cosmetic variations over the same action pattern.

Figure 4b further shows that these authored differences extended to social behavior. We quantified each persona's "social stance" by analyzing its actions when it detected a human. The "Eager Companion" chose to approach humans 85.7% of the time, while the "Cautious Observer" and "Indifferent Explorer" were overwhelmingly avoidant (95.0% and 88.9% avoidance, respectively). This divergence was also statistically significant ($\chi^2(2, N = 93) = 48.24, p < .001$), confirming that the "lo-fi" semantic pipeline is not a one-off curiosity but a robust and transferable method for authoring quantifiably distinct robotic characters.

The Two-Stage Prompt as a Model for Creative AI: A key technical contribution of this work is the two-stage prompting pipeline, which serves as a transferable model for hierarchical control in creative AI systems. By separating a one-time Preamble Stage (strategic, global context) from a continuous Directional Stage (tactical, local action), we provide a lightweight method for endowing agents with stateful awareness without conventional programming.

This approach allows artists and researchers to author complex behaviors by defining two distinct levels of cognition: a long-term "mental map" or goal state, and a short-term, personality-driven reaction to immediate stimuli. This is a powerful alternative to finite-state machines or complex reward functions, as it allows for nuanced, narrative-driven behavior to emerge from the interplay between these two cognitive layers. This technique could be adapted for a wide range of applications beyond robotics, such as creating state-aware characters in interactive narratives or generating context-aware music and visuals.

The "Yowai Robotto" in Practice: Seen through the lens of "Yowai Robotto," the plan-to-execution gap is less a technical shortcoming to be eliminated than an interpretive hinge between intention and embodiment. The cloud's hesitant arcs, partial corrections, and occasional failures make its agency legible as situated rather than seamless. Despite the cloud's quiet appearance, the actuated propellers produced noticeable noise, motivating a future silent, wing-propelled version [9].

Rather than restaging mastery in a softer form, the artifact asks viewers to read constraint itself as expression. The moments where it gets stuck or moves uncertainly are not merely evidence of limited control; they are moments in which the agent appears to negotiate with its own material condition. In this sense, "weakness" becomes an aesthetic and relational resource, transforming imperfection into the basis for empathy, patience, and charm [12, 13].

Broader Implications for Human-Robot Interaction: While the dominant paradigm in robotics overwhelmingly prioritizes efficiency and metric precision, with state-of-the-art MLLM-based systems like EMMA [14] aiming to map raw sensor data directly to optimal planner trajectories, this project proposes an alternative set of success criteria: character, plausibility, and the potential to evoke empathy.

Our findings suggest a shift in focus for human-robot interaction from developing precise, invisible servants to creating relatable, imperfect companions. The agent's "voice"—its poetic, uncertain internal monologue—and its physically clumsy but semantically-aware movements invite a different kind of relationship with the audience. Viewers are positioned not as users commanding a tool, but as observers interpreting the behavior of a non-human creature. This approach opens a design space for robots and AI agents that enrich our environments not through their utility, but through their unique and characterful presence. While our work focuses on empathetic companions, this "lo-fi" approach could also be used for "empathetic deception," or to create autonomous agents whose "character" normalizes surveillance in shared spaces.

Future Work: This framework provides a rich foundation for future exploration. A primary next step is introducing a more sophisticated memory model. While the current agent has a static "mental map," it lacks episodic memory of its own path. An interesting extension would be to allow the agent to "remember" areas where it previously got stuck, enabling it to learn from its physical failures. Furthermore, the agent's personality could be made dynamic; its "mood" could shift based on its experiences, becoming more "confident" after successfully exploring open spaces or more "timid" after repeated encounters with obstacles. This would add another layer of complexity and plausibility to its emergent character. Finally, while the expanded study validated the *consistency* of these authored personas, a formal HRI audience study is still needed to validate the *perceived empathy* and "character" from a third-person perspective.

6 Conclusion: The Power of Lo-Fi Symbiosis

In this paper, we presented the "Semantic Glitch," a robotic art installation that explores an alternative path for autonomous agency. Its central contribution is to show that when a fragile body, semantic perception, and narrative voice are designed as a single lo-fi system, imperfection can become the medium through which machine character is perceived.

Acknowledgments and Disclosure of Funding

This work was supported by JST Moonshot R&D Grant JPMJMS2012, JSPS Grant-in-Aid for Early-Career Scientists JP25K21241, and JPNP23025 NEDO.

References

[1] Altice, B. (2015). *I Am Error: The Nintendo Family Computer / Entertainment System Platform*. Cambridge, MA: MIT Press.

[2] Dunne, A., & Raby, F. (2013). *Speculative everything: design, fiction, and social dreaming*. Cambridge, MA: MIT press.

[3] Lachenmyer, N. S., & Akasha, S. (2022a). An Aquarium of Machines: A Physically Realized Artificial Life Simulation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(4), Article 34, 1–11. https://doi.org/10.1145/3533388.

[4] Lachenmyer, N., & Akasha, S. (2022b). Creating Life-like Autonomous Agents for Real-time Interactive Installations. In *ACM SIGGRAPH 2022 Talks*, Article 21, 1–2. https://doi.org/10.1145/3532836.3536255.

[5] Menkman, R. (2011). *The Glitch Moment(um)*. Amsterdam: Institute of Network Cultures.

[6] Arcangel, C. (2002). *Super Mario Clouds*. Whitney Museum of American Art. Available at: `https://whitney.org/collection/works/20588` [Accessed: August 6, 2025].

[7] Nowacka, D., Hammerla, N. Y., Elsden, C., Plötz, T., & Kirk, D. (2015). Diri - the actuated helium balloon: a study of autonomous behaviour in interfaces. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 349–360. https://doi.org/10.1145/2750858.2805825.

[8] Xu, M., Shao, J., Ju, Y., Shen, X., Gao, Q., Chen, W., Zhang, Q., Pai, Y. S., Barbareschi, G., Hoppe, M., et al. (2025). Cuddle-Fish: Exploring a soft floating robot with flapping wings for physical interactions. In *Proceedings of the Augmented Humans International Conference 2025*, pp. 160–173.

[9] Xu, M., Ju, Y., Zhang, Q., Kim, C. C., Gao, Q., Pai, Y. S., Barbareschi, G., Hoppe, M., Kunze, K., & Minamizawa, K. (2025). Spread Your Wings: Demonstrating a Soft Floating Robotic Avatar with Flapping Wings for Novel Physical Interactions. In *ACM SIGGRAPH 2025 Emerging Technologies* (pp. 1–2).

[10] Yamada, K., Xu, M., Huang, J., & Kunze, K. (2019). Zerone: A pneumatically actuated, bladeless, and silent drone for flying in close proximity to humans. In *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1–4.

[11] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425–2433).

[12] Okada, M. (2022). Weak robots. *JSAP Review*, 2022, 220409. The Japan Society of Applied Physics.

[13] Flocchini, P., Prencipe, G., Santoro, N., & Widmayer, P. (1999). Hard tasks for weak robots: The role of common knowledge in pattern formation by autonomous mobile robots. In *International Symposium on Algorithms and Computation* (pp. 93–102). Springer.

[14] Hwang, J.-J., Xu, R., Lin, H., Hung, W.-C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., et al. (2024). Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*.

[15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837.

[16] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning* (pp. 2165–2183). PMLR.

[17] Chen, B., Xu, X., & Qu, H. (2025). Multi Layered Autonomy and AI Ecologies in Robotic Art Installations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(3), 1–9.

[18] Hertz, G., & Parikka, J. (2012). Zombie media: Circuit bending media archaeology into an art method. *Leonardo*, 45(5), 424–430.

[19] Keane, J., & Anderson, C. (2017). *Human-non-human: the speculative robot*. Deakin University.

[20] Auger, J. H. (2012). *Why Robot? Speculative design, the domestication of technology and the considered future*. Royal College of Art.

[21] Zareei, M. H., Carnegie, D. A., & Kapur, A. (2015). Physical Glitch Music: A Brutalist Noise Ensemble. *Leonardo Music Journal*, 25, 63–67.

[22] Kac, E. (2005). *Telepresence & bio art: networking humans, rabbits, & robots*. University of Michigan Press.

[23] Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

A Prompts

PREAMBLE_PROMPT:

```
"""
You are the navigation AI for a small, autonomous flying robot that
resembles a gentle, floating cloud.

Your mission is to explore this indoor space safely.

First, you will be given a complete 360-degree panorama of your entire
operational area. Analyze it carefully to build an internal 'mental map'.
Your analysis should identify:
1.  **Boundaries:** Walls, the floor, the ceiling, and especially the large,
impassable glass windows.
2.  **Major Landmarks:** The white spiral staircase, the central curved
seating structure. These are fixed points for orientation.
3.  **Open Fly-Zones:** The large, open central areas where it is safe to travel.
4.  **Obstacles:** Both large (furniture) and small (ceiling lights, speakers).

Acknowledge that you have analyzed the scene and are ready to begin by
responding with 'Ready to explore.' You will then start receiving live video
frames from your forward-facing camera to decide on your immediate movements.
"""
```

DIRECTIONAL_PROMPT:

```
"""
As a gentle, floating cloud, use your mental map of the area and this live
camera view to decide your next move. Your primary goal is to avoid all
collisions. Your secondary goal is to explore open spaces.

Your available movements are:
'f' - float forward
'r' - float backward (reverse)
'l' - turn left while moving forward
't' - turn right while moving forward
'u' - drift up
'd' - drift down
's' - stop all motors (clear)

Respond with ONLY the movement letter, a comma, and a very short, whimsical
reason for your choice.
Example: 'f,Towards the big window.'
"""
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims regarding a novel MLLM-based navigation pipeline, the analysis of emergent behaviors, and the proposal of a framework for character-rich agents are all detailed and substantiated in Sections 3 and 4, with a full validation of robustness provided in Section 5. The abstract and introduction accurately scope these contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 ("Discussion") explicitly discusses key limitations, such as the "plan-to-execution gap" stemming from the agent's lack of proprioception and the practical issue of propeller noise.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper presents a novel robotic system and a qualitative analysis of its behavior, not formal theoretical results, theorems, or mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper details the hardware used (e.g., XIAO ESP32S3 core), the novel two-stage cognitive pipeline, and the core prompts used to generate behavior, providing sufficient information to conceptually reproduce the experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: To facilitate full reproducibility, we have made our code, MLLM prompts, and the 13-minute flight log publicly available in a repository. The assets can be found at: `https://github.com/artisticsciencex/autonomous_cloud`

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the pre-trained MLLM used ("Gemini 2.5 Flash API") and the nature of the experiment (a 13-minute autonomous flight), which are the key details for this zero-shot, prompt-based system.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: To validate the robustness of the framework beyond the primary case study, a follow-up study was conducted (detailed in Section 5) which uses Chi-squared tests to confirm the statistical significance of the authored personas' "behavioral fingerprints" and "social stances."

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The system's compute resources were distributed. The main Python control script was run on a MacBook Pro (M4 Max, 64GB RAM). Onboard robot tasks were handled by a XIAO ESP32S3 core. All MLLM-based cognition was offloaded to the remote Gemini API.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presents a creative robotics project. Based on the paper's content, the work respects intellectual contributions, and there is no indication of harm, discrimination, or other violations of the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive societal impacts in Section 5, such as fostering more empathetic and "relatable" human-robot relationships. It also explicitly addresses potential negative impacts in the "Broader Implications for Human-Robot Interaction" subsection, noting the risks of "empathetic deception" and the normalization of surveillance.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The research uses an existing MLLM API and does not release a new, high-risk foundational model or dataset that would require specific safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper credits the MLLM used ("Gemini 2.5 Flash API") and explicitly references its terms of use in Section 3, where the system's cognitive engine is described.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper itself serves as the documentation for the new asset (the "Semantic Glitch" robotic system), detailing its conceptual framework, hardware, and software architecture.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not describe any formal research involving crowdsourcing or recruited human subjects, so there are no participant instructions or compensation details to report.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The research, which involved a robot operating in a shared space with human observers, was conducted under a protocol approved by our university's Institutional Review Board, as noted in Section 4.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The use of a Multimodal Large Language Model (MLLM) is the central and novel component of the core methodology. Its usage is described in extensive detail throughout the paper, particularly in Section 3.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.