Research Statement:
Towards Coherent Embodied Conversational Agent

Jaewoo Ahn
Department of Computer Science and Engineering
Seoul National University
jaewoo.ahn@vision.snu.ac.kr

1 Introduction

One of the ultimate goals in artificial intelligence (AI) is to create a *cognitive partner* that can interact and collaborate with humans through natural conversation, moving beyond mere command-execution tools. At the heart of this vision lies the **Embodied Conversational Agent (ECA)** [1], an intelligent system with a physical or virtual form that allows it to perceive, act, and communicate within its environment. While the advent of foundation models, such as Large Language Models (LLMs), has revolutionized the language capabilities of AI, their knowledge remains fundamentally disconnected from the physical world. The core challenge in realizing a truly capable ECA is to bridge this gap between abstract language understanding and concrete physical reality.

My research addresses this grand challenge by focusing on a central theme: **Coherence**. To become a reliable partner, a successful ECA must maintain coherence across its dialogue, perception, and actions. I have systematically pursued this goal by establishing three fundamental research pillars. The first is **Conversation & Interaction**, building the agent's ability to engage in dialogue that is grounded in real-world information and a consistent identity. The second is **Robust Multimodal Perception**, ensuring the agent can understand the physical and conceptual world without hallucinations or distortions. The final pillar is **Embodied Decision-Making**, which enables the agent to formulate and execute intelligent plans to achieve long-term goals based on its understanding.

This research statement details my journey toward this vision. Section 2 introduces my foundational work in establishing the basis of conversational agents that are natural and persona-driven. Section 3 describes my efforts to build a robust multimodal perception system that mitigates hallucinations and understands compositional concepts. Section 4 discusses the benchmarks and novel agentic frameworks I developed for solving long-horizon tasks in complex simulated environments. Finally, Section 5 outlines my future research directions, building upon these contributions to realize a true cognitive partner.

2 Conversation & Interaction

The foundation of an effective ECA lies in its ability to conduct natural and meaningful interactions with humans. My research began by building the core components necessary for an agent to move beyond simplistic responses and engage in coherent conversations grounded in information, identity, context, and time. These foundational studies focused on establishing informational, identity, contextual, and temporal coherence, creating a solid bedrock for the perception and decision-making research that followed.

Knowledge-Grounded Dialogue for Informational Coherence A primary challenge for early dialogue systems was their tendency to produce uninformative and repetitive responses. To address this, I focused on enabling agents to dynamically select and utilize external knowledge in multi-turn conversations. My co-authored work, SKT (Sequential Knowledge Transformer) [2], proposed a novel sequential latent variable model to track the conversational flow and select the most relevant knowledge at each turn. This approach achieved state-of-the-art performance on large-scale benchmarks like Wizard of Wikipedia, demonstrating that an agent could maintain informational coherence and engage in more substantive dialogue.

Multimodal & Persona-Grounded Memory for Identity and Contextual Coherence To create more human-like agents, I extended my research to persona-based dialogue, allowing agents to converse based on a unique personality and memory. Pushing beyond the limits of text-only personas, my work on MPChat [3] introduced the first 'multimodal persona' by combining images and text to represent a user's episodic memories. This allows an agent to establish identity coherence and enables richer, more personalized interactions. Furthermore, to ensure agents can grasp the full context of complex online discussions, we developed the mRedditSum dataset [4] and a novel methodology to summarize long, multimodal threads. This work secured an agent's ability to maintain contextual coherence by identifying key information in complex, multi-speaker conversations.

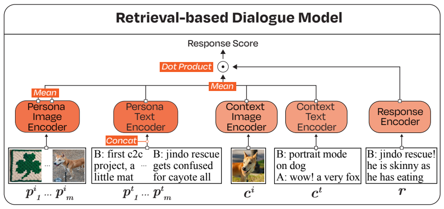


Figure 1: I proposed a dialogue model for multimodal persona-grounded conversations [3].

Ensuring Temporal Coherence in LLMs The advent of LLMs, while powerful, introduced a new challenge: "Point-in-Time Character Hallucination," where a role-playing agent situated at a specific point in a narrative erroneously reveals knowledge of future events. To address this, I created TimeChara [5], the first benchmark to rigorously measure this phenomenon. I also proposed Narrative-Experts, a method that enhances spatiotemporal consistency by decomposing the reasoning process. This research pioneered the study of temporal coherence, ensuring that LLM-based agents can stay faithfully immersed in a specific role and timeline.

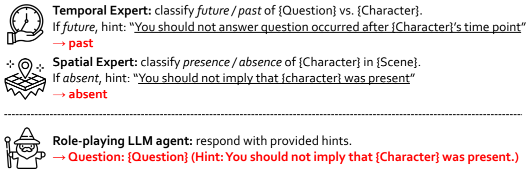

Temporal Expert: classify future / past of {Question} vs. {Character}.
If future, hint: "You should not answer question occurred after {Character}'s time point"
→ past
Spatial Expert: classify presence / absence of {Character} in {Scene}.
If absent, hint: "You should not imply that {character} was present"
→ absent
Role-playing LLM agent: respond with provided hints.
→ Question: {Question} (Hint: You should not imply that {Character} was present.)

Figure 2: I proposed Narrative-Experts method to solve point-in-time character hallucination by decomposing LLM's reasoning steps [5].

3 Robust Multimodal Perception

For an ECA to be coherent, its ability to converse must be grounded in an accurate and robust understanding of the world. If an agent misperceives its environment or hallucinates information that isn't there, any subsequent decision-making and action become unreliable. My research has therefore focused on building the agent's capacity to perceive the world—both physical and conceptual—with unwavering consistency. I have pursued this by starting with the foundations of conceptual coherence, then moving to diagnose fundamental vulnerabilities in multimodal models, and finally developing practical solutions to mitigate hallucinations.

Conceptual Coherence as a Foundation for Perception The deepest foundation of robust perception is conceptual coherence—the ability to understand how concepts combine and how their properties evolve. An agent must know, for instance, that a "peeled apple" is no longer "red." In my co-authored work on CCPT (Conceptual Combination with Property Type) [6], we introduced the first benchmark to evaluate this fundamental reasoning skill in LLMs. This research revealed that LLMs struggle with generating and understanding emergent properties, and we proposed a novel method inspired by the 'spreading activation' model from cognitive psychology to improve this capability. This work lays the groundwork for the commonsense reasoning that is essential for an agent to correctly interpret its physical world.

Probing Compositional Vulnerabilities in Multimodal Representations Building on this conceptual foundation, I analyzed the compositional vulnerabilities of leading multimodal perception models. Models like CLIP often fail to distinguish between "a baby is on a bed" and "a bed is on a baby," revealing a critical weakness in understanding relationships between objects. To address this, I proposed the MAC (Multimodal Adversarial Compositionality) benchmark [7]. MAC utilizes LLMs to generate adversarial text that deceives perception models across various modalities, including images, video, and audio. It introduces a novel evaluation framework that measures not only the attack success

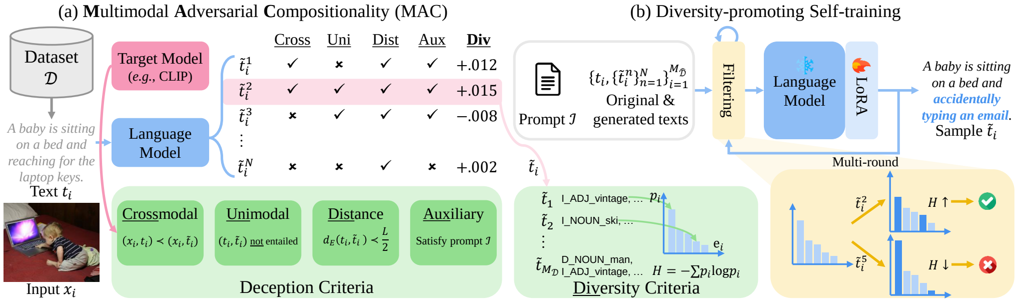

(a) Multimodal Adversarial Compositionality (MAC)
Dataset $\mathcal{D}$
Target Model (e.g., CLIP)
A baby is sitting on a bed and reaching for the laptop keys.
Text t_i
Input x_i
Language Model
Cross
Uni
Dist
Aux
Div
$\tilde{t}_i^1$ ✓ × ✓ ✓ +.012
$\tilde{t}_i^2$ ✓ ✓ ✓ ✓ +.015
$\tilde{t}_i^3$ × ✓ ✓ ✓ −.008
$\tilde{t}_i^N$ × × ✓ × +.002
Crossmodal
$(x_i, t_i) \prec (x_i, \tilde{t}_i)$
Unimodal
$(t_i, \tilde{t}_i)$ not entailed
Distance
$d_E(t_i, \tilde{t}_i) \prec \frac{L}{2}$
Auxiliary
Satisfy prompt $\mathcal{J}$
Deception Criteria
(b) Diversity-promoting Self-training
Prompt $\mathcal{J}$
$\{t_i, \{\tilde{t}_i^n\}_{n=1}^N\}_{i=1}^{M_\mathcal{D}}$
Original & generated texts
Filtering
Language Model
LoRA
A baby is sitting on a bed and accidentally typing an email.
Sample $\tilde{t}_i$
Multi-round
$\tilde{t}_i$
$\tilde{t}_1$ I_ADJ_vintage, …
$\tilde{t}_2$ I_NOUN_ski, …
$\tilde{t}_{M_\mathcal{D}}$ D_NOUN_man, I_ADJ_vintage, …
p_i
e_i
$H = -\sum p_i \log p_i$
Diversity Criteria
$\tilde{t}_i^2$
$\tilde{t}_i^5$
$H \uparrow$
$H \downarrow$

Figure 3: I proposed (a) MAC benchmark, and (b) diversity-promoting self-training method to enhance attack success rate as well as attack diversity [7].

rate but also the 'diversity' of the attack methods. Furthermore, 'diversity-promoting self-training' method proved that even small LLMs (~8B) can be used to effectively uncover vulnerabilities in multimodal models, paving the way for developing more robust perception systems.

Mitigating Hallucination in Structured Visual Data Beyond analyzing vulnerabilities, I have worked to solve the hallucination problem in a specific domain: structured visual data like charts. Vision-language models (VLMs) often hallucinate when describing charts, largely due to training datasets containing extraneous information. To solve this, We developed ChartCap [8], a large-scale dataset of 565K real-world charts with captions that are dense, type-specific, and, crucially, free from external information. We also introduced an innovative, reference-free evaluation metric, the Visual Consistency Score (VCS), which reconstructs a chart from a generated caption and measures its visual similarity to the original. Models trained on ChartCap produced captions that were more accurate and hallucination-free than even those from powerful proprietary models like Claude 3.5 Sonnet or those written by humans, providing a practical solution for reliable data interpretation.

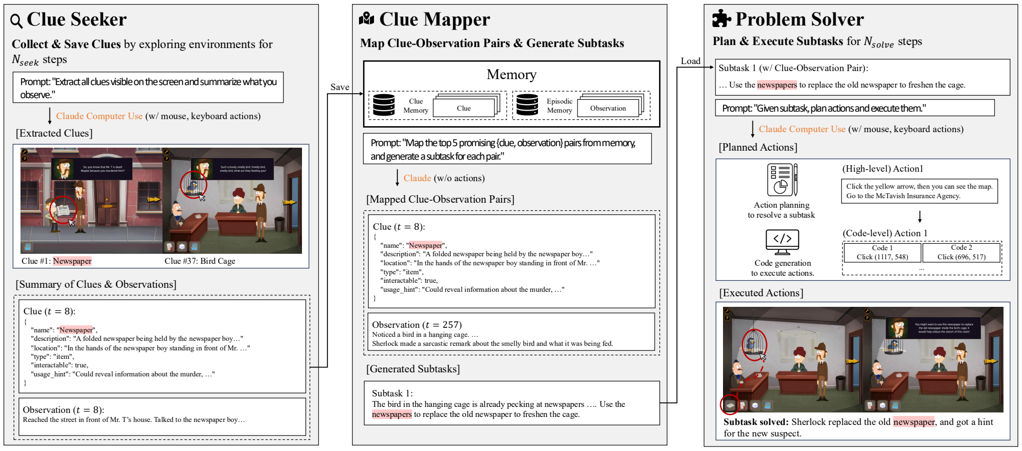
4 Embodied Decision-Making

An agent that can converse and perceive is still incomplete without the ability to act. For an ECA to be truly valuable, it must translate its understanding into intelligent Decision-Making to achieve long-term goals within complex, dynamic environments. My research has culminated in this area, focusing on creating agents that can strategically plan and execute actions. To this end, I have developed a foundational benchmark to evaluate a wide range of agentic capabilities and proposed a novel framework that addresses the critical challenge of long-term, memory-based problem-solving.

A Foundational Benchmark for Diverse Gameplay To evaluate and advance the decision-making capabilities of LLM agents, a realistic and comprehensive testbed is essential. Addressing this need, we developed Orak [9], a foundational benchmark built on 12 real-world, commercially successful video games across six major genres. Orak is more than a collection of games; it is a complete platform that (1) enables a holistic evaluation of diverse LLM capabilities, from reaction time to strategic planning, (2) facilitates in-depth studies on the impact of agentic modules like self-reflection and planning, and (3) provides the first fine-tuning dataset of expert gameplay trajectories to help specialize general LLMs into capable gaming agents. Through its plug-and-play interface powered by the Model Context Protocol (MCP), Orak establishes a stable and consistent environment for evaluating the rapidly evolving landscape of LLM agents.

Solving Full Story Arcs via Long-Term Memory One of the most significant hurdles in long-horizon tasks is the "observation-behavior gap": the challenge of remembering a clue from an early stage (observation) and applying it much later to solve a problem (behavior), which demands long-term memory and high-level reasoning. To tackle this problem head-on, I created the FlashAdventure benchmark [10], a collection of 34 complete adventure games that require agents to solve full story arcs from start to finish. To overcome the limitations of manual evaluation, I also proposed CUA-as-a-Judge, an automated system that verifies gameplay progress.

Crucially, to bridge the observation-behavior gap, I developed COAST (Clue-Oriented Agent for Sequential Tasks), a novel agentic framework. COAST operates on a "Seek-Map-Solve" cycle: it



Clue Seeker
Collect & Save Clues by exploring environments for N_{seek} steps

Prompt: "Extract all clues visible on the screen and summarize what you observe."

Claude Computer Use (w/ mouse, keyboard actions)
[Extracted Clues]
So, you know that Mr. T is dead! Maybe because you murdered him?!
Such a lovely smelly bird. Smelly bird, smelly bird, what are they feeding you?
EXTRA!
Clue #1: Newspaper
Clue #37: Bird Cage
[Summary of Clues & Observations]

Clue ($t = 8$):

```
{
  "name": "Newspaper",
  "description": "A folded newspaper being held by the newspaper boy…"
  "location": "In the hands of the newspaper boy standing in front of Mr. …"
  "type": "item",
  "interactable": true,
  "usage_hint": "Could reveal information about the murder, …"
}
```

Observation ($t = 8$):
Reached the street in front of Mr. T's house. Talked to the newspaper boy…

Prompt: "Map the top 5 promising {clue, observation} pairs from memory, and generate a subtask for each pair."

Clue ($t = 8$):

```
{
  "name": "Newspaper",
  "description": "A folded newspaper being held by the newspaper boy…"
  "location": "In the hands of the newspaper boy standing in front of Mr. …"
  "type": "item",
  "interactable": true,
  "usage_hint": "Could reveal information about the murder, …"
}
```

Observation ($t = 257$)
Noticed a bird in a hanging cage. …
Sherlock made a sarcastic remark about the smelly bird and what it was being fed.

Subtask 1:
The bird in the hanging cage is already pecking at newspapers …. Use the newspapers to replace the old newspaper to freshen the cage.

Subtask 1 (w/ Clue-Observation Pair):
… Use the newspapers to replace the old newspaper to freshen the cage.

Prompt: "Given subtask, plan actions and execute them."

Claude Computer Use (w/ mouse, keyboard actions)
[Planned Actions]
Action planning to resolve a subtask
Code generation to execute actions.
(High-level) Action1

Click the yellow arrow, then you can see the map. Go to the McTavish Insurance Agency.

(Code-level) Action 1
Code 1
Click (1117, 548)
Code 2
Click (696, 517)
...
[Executed Actions]
You might want to use this newspaper to replace the old newspaper inside the bird's cage. It would help reduce the stench of this room!

Subtask solved: Sherlock replaced the old newspaper, and got a hint for the new suspect.

Figure 4: I proposed COAST framework with Seek-Map-Solve cycle to solve long-term observation-behavior gap [10].

(1) proactively explores to gather clues into a long-term memory, (2) maps connections between clues and past observations to generate meaningful subtasks, and then (3) executes those subtasks. In our experiments, COAST significantly outperformed state-of-the-art GUI agents in both story completion and milestone achievement rates, proving that memory-driven reasoning is essential for solving complex, long-horizon problems.

5 Future Directions

My research has established the three pillars of an ECA—Conversation, Perception, and Decision-Making—under the unifying theme of Coherence. To this end, I outline two groups of future directions: (1) advancing each pillar in depth, and (2) integrating pillars for richer collaboration.

5.1 Deepening Individual Pillars

Enhancing Long-Horizon Task Performance via Reinforcement Learning Orak [9] and FlashAdventure [10] revealed that agents still struggle with strategic consistency in long-horizon tasks, especially under sparse rewards. I will address this by designing an RL framework that generates dense rewards from auxiliary signals such as sub-goal completion or conversational feedback.

Strengthening Robust Perception and Reasoning My work on MAC [7] showed that current models remain vulnerable in compositional reasoning even for well-studied modalities such as vision and audio. I plan to advance this line of research in two directions: (1) strengthening perception and reasoning in these existing modalities by developing more compositional and causally grounded models, and (2) extending to new sensory modalities such as tactile and thermal signals.

5.2 Integrating Across Pillars

Embodied Theory of Mind for Multi-Agent Interaction Beyond task execution, effective interaction requires understanding others' mental states. Building on TimeChara [5], I will develop agents with an Embodied Theory of Mind (ToM) that infer intentions, beliefs, and emotions from multimodal cues (language, gaze, actions). This line of work integrates the pillars of conversation and decision-making.

Bridging the Sim-to-Real Gap for Physical Grounding While the frameworks developed in Orak [9] and FlashAdventure [10] have proven effective in simulation, the ultimate testbed for an ECA is the physical world. My final research thrust will be a Sim-to-Real initiative to transfer these advances to a physical robot platform. This direction naturally integrates the decision-making and robust perception pillars, ensuring that agents can reason and act coherently in complex, dynamic environments beyond simulation.

References

[1] Vardhan Dongre, Dilek Hakkani-Tür, and Gokhan Tur. Embodied conversational agents and the promise of foundation models. `https://tinyurl.com/embodiedconvai`, 2025. [Accessed 12-02-2025].

[2] Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. Sequential latent knowledge selection for knowledge-grounded dialogue. In *ICLR*, 2020.

[3] Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. MPCHAT: Towards multimodal persona-grounded conversation. In *ACL*, 2023.

[4] Keighley Overbay, Jaewoo Ahn*, Fatemeh Pesaran zadeh*, Joonsuk Park, and Gunhee Kim. mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images. In *EMNLP*, 2023.

[5] Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *ACL Findings*, 2024.

[6] Seokwon Song*, Taehyun Lee*, Jaewoo Ahn, Jae Hyuk Sung, and Gunhee Kim. Is a peeled apple still red? evaluating LLMs' ability for conceptual combination with property type. In *NAACL*, 2025.

[7] Jaewoo Ahn*, Heeseung Yun*, Dayoon Ko, and Gunhee Kim. Can LLMs deceive CLIP? benchmarking adversarial compositionality of pre-trained multimodal representation via text updates. In *ACL*, 2025.

[8] Junyoung Lim, Jaewoo Ahn, and Gunhee Kim. ChartCap: Mitigating hallucination of dense chart captioning. In *ICCV*, 2025.

[9] Dongmin Park*, Minkyu Kim*, Beongjun Choi*, Junhyuck Kim, Keon Lee, Jonghyun Lee, Inkyu Park, Byeong-Uk Lee, Jaeyoung Hwang, Jaewoo Ahn, Ameya S. Mahabaleshwarkar, Bilal Kartal, Pritam Biswas, Yoshi Suhara, Kangwook Lee, and Jaewoong Cho. Orak: A foundational benchmark for training and evaluating llm agents on diverse video games. *arXiv preprint arXiv:2506.03610*, 2025.

[10] Jaewoo Ahn*, Junseo Kim*, Heeseung Yun, Jaehyeon Son, Dongmin Park, Jaewoong Cho, and Gunhee Kim. FlashAdventure: A benchmark for gui agents solving full story arcs in diverse adventure games. In *EMNLP*, 2025.