APUT: Large Language Models as Cross-Modal Consistency Reasoning Engines for Egocentric State Estimation

Haochen Huang*

Beijing 101 Middle School, Beijing, China, 100091

*Corresponding Author. Email: cldcloud@outlook.com

Abstract

Current multimodal large language models (LLMs) excel at passive representation alignment. However, in highly dynamic, partially observable environments (e.g., egocentric vision with severe occlusion and the "cocktail party" effect), standard models suffer from representational collapse. Unobserved targets are frequently dropped from working memory or their states are hallucinated. To address this, the Active-Perception Unified Transformer (APUT) is proposed. Crucially, APUT is not designed as an improved multimodal tracker, but rather as a reasoning-based state estimator. Identity tracking is treated as a dynamic constraint satisfaction problem (CSP). By modeling a structured state matrix through a dedicated encoder, the LLM operates over explicit tokenized world states. Implicit latent tracking collapses under contradictory cross-modal evidence because state representations are entangled, whereas explicit constraint graphs allow for discrete hypothesis elimination. Furthermore, a partial delta-update mechanism and a multiphase training pipeline, ranging from teacher-forced alignment to critic-guided hypothesis ranking, are introduced. Through this architecture, it is demonstrated how robust object permanence emerges from internal logical coherence under supervised structured learning, providing a novel paradigm for embodied state estimation.

Key words

cross-modal consistency, egocentric state estimation, active perception, object permanence, constraint satisfaction

1. Introduction

Maintaining the identity and state of multiple entities in unconstrained egocentric environments—characterized by rapid ego-motion, prolonged occlusion, and overlapping acoustics—is a central challenge for embodied artificial intelligence [1].

A superficial reading might mischaracterize APUT

as a mere integration of object tracking, audio separation, and memory regularization [2]. However, such traditional pipelines passively map sensory inputs to output tracks. When sensory evidence vanishes (e.g., severe occlusion), these systems collapse. Implicit latent tracking collapses under contradictory cross-modal evidence because state representations are entangled, whereas explicit constraint graphs allow discrete hypothesis elimination. It is argued that human cognition maintains object permanence not by passive tracking, but by treating the environment as a continuous logic puzzle. APUT shifts the paradigm from "how to fuse modalities" to "how to estimate and maintain global state consistency." In this framework, the LLM acts as a global consistency solver. When inputs conflict, the LLM evaluates multiple hypotheses across interacting object files, enforces mutual exclusivity, and actively generates audio masks to isolate evidence, logically resolving the global state matrix.

2. Literature review

Multimodal alignment and robustness. Foundational encoders like Video Prism [3] and early fusion models like Chameleon [4] focus on robust, dense feature representations. However, these architectures treat video as a feed-forward sequence, lacking the internal recurrent states necessary to reason about unobserved entities over long temporal horizons. The reliance on passive observation limits their applicability in scenarios where continuous tracking is required despite frequent occlusions. Recent advancements have attempted to bridge this gap by introducing temporal attention mechanisms, yet they often fall short when faced with contradictory multimodal inputs. The inherent difficulty lies in their inability to actively seek clarifying information or maintain a structured memory of entities that temporarily disappear from the sensory field. This passive alignment paradigm is insufficient for complex egocentric tasks where object permanence is critical. Consequently, there is a pressing need for architectures that can actively manage and reason over explicit state representations rather than relying solely on dense feature matching across frames.

Tool-augmented and ReAct agents. Frameworks that integrate LLMs with external APIs (e.g., zero-shot audio separators like AudioSep [5]) view the LLM as a scheduler. While APUT also triggers audio separation, its objective is fundamentally different: separation is not the end goal, but a hypothesis-testing mechanism triggered internally to prune conflicting branches in a global consistency graph. Existing ReAct-style agents often employ external tools in a linear, task-oriented manner, lacking the deep integration required for continuous state estimation. They treat tools as black boxes for information retrieval rather than as active components in a dynamic reasoning process. APUT, in contrast, tightly couples tool usage with internal state updates, allowing the LLM to actively resolve ambiguities by selectively querying the environment. This paradigm shift from simple tool invocation to active hypothesis verification is crucial for maintaining a coherent understanding of dynamic scenes. The integration of such active perception mechanisms represents a significant departure from traditional agent architectures and opens new avenues for robust multimodal reasoning in unconstrained settings.

Object-centric representation. Models like Slot-VLM [6] maintain implicit latent slots. Unlike black-box slots, the episodic object files in APUT are explicit state variables subjected to logical constraint checking by the LLM, enabling zero-shot conflict resolution based on physical commonsense. Implicit representations often struggle to disentangle complex interactions between multiple entities, leading to the hallucination of states or the loss of object identities. By making these state variables explicit and tokenized, APUT allows the LLM to directly manipulate and reason over them. This explicit modeling facilitates the application of logical constraints, such as spatial exclusivity and temporal continuity, which are essential for robust tracking. Furthermore, explicit representations enhance interpretability, as the internal reasoning process of the model can be directly observed and analyzed. The transition from implicit latent spaces to explicit structured representations is a key innovation in the pursuit of reliable embodied perception, and it aligns closely with cognitive science theories of object file maintenance [7].

3. Methodology: the LLM as a consistency solver

3.1 Structured state matrix and dedicated encoder

Instead of treating profiles as implicit hidden vectors, APUT models the environment as a structured constraint graph $G_t = (V_t, E_t)$. The proposed approach models the structured state matrix through a dedicated encoder, allowing the LLM to operate over tokenized representations of explicit world states. Each node $v^{(i)} \in V_t$ represents an episodic object file for target i, formatted as shown in equation (1).

$$v_t^{(i)} = [ID^{(i)}, \mathcal{S}_{anchors}^{(i)}, \mathcal{P}_{spatial}^{(i)}, Note_{ep}^{(i)}] \quad (1)$$

Here, $\mathcal{S}_{anchors}$ contains high-dimensional voiceprint and visual signatures. The edges E_t represent physical and acoustic constraints (e.g., spatial exclusivity: $\mathcal{P}_{spatial}^{(i)} \neq \mathcal{P}_{spatial}^{(j)}$).

3.2 Interleaved generation and active verification

When constraint propagation yields multiple plausible hypotheses, APUT initiates an active verification loop. The LLM pauses standard text generation and emits discrete audio mask tokens. These tokens direct a deterministic bottom-up decoder to extract purified acoustic signals, which are re-injected into the LLM as new explicit evidence to prune the constraint graph. The overall workflow is illustrated in Figure 1.

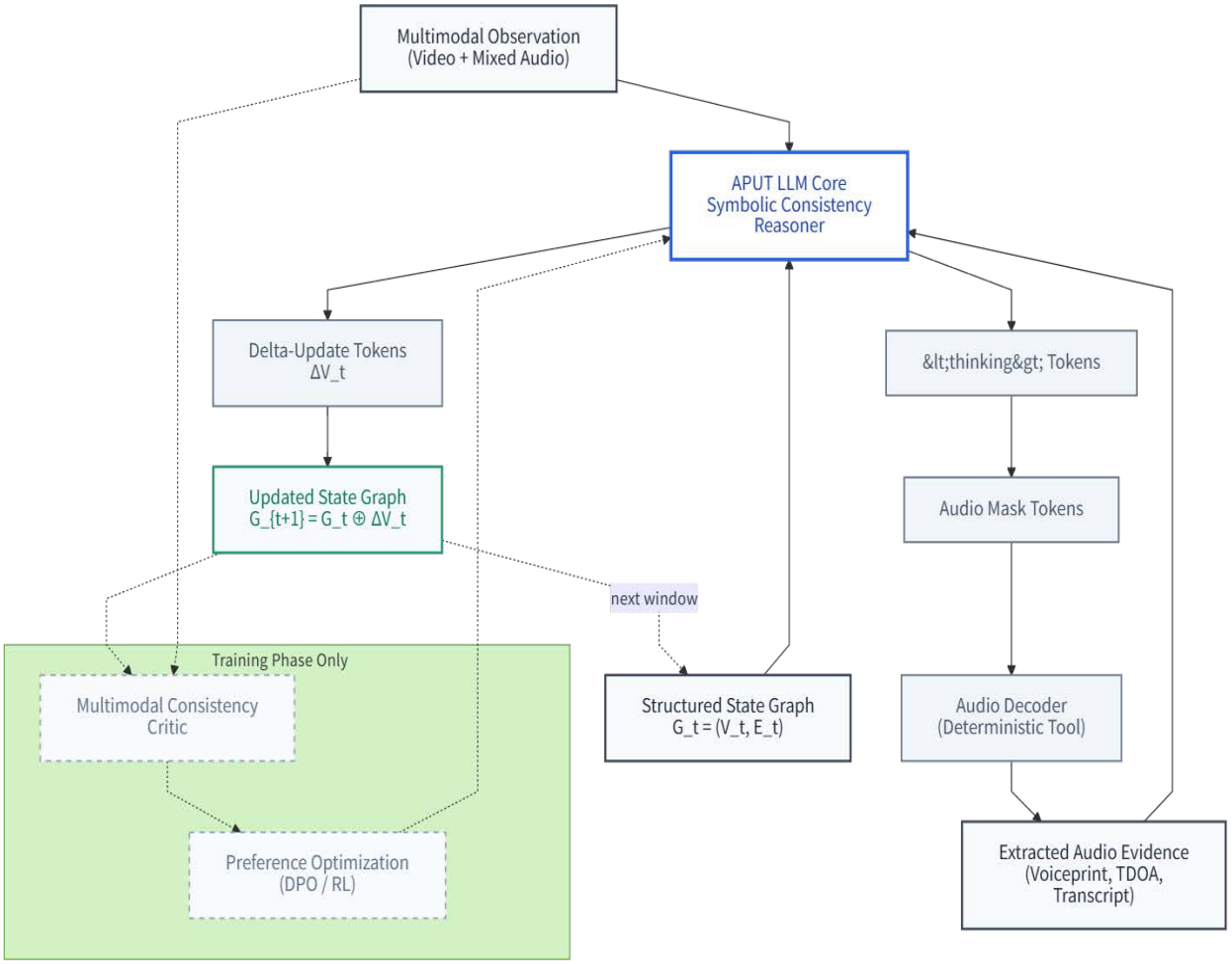

Figure 1: The interleaved generation workflow. The LLM pauses text generation to output audio mask tokens, triggering an external decoder. The extracted audio properties are re-injected as observations to resolve cross-moral ambiguity

3.3 Incremental inference via delta-updates

Instead of regenerating the entire global state matrix at every timestep, APUT employs a partial update mechanism. The LLM generates delta-updates (ΔV_t) that modify only the affected slots within the constraint graph as formulated in equation (2).

$$G_{t+1} = G_t \oplus \Delta V_t \tag{2}$$

This design stabilizes training, preserves persistent context without catastrophic forgetting, and aligns naturally with the intended "Sudoku-style" reasoning process, in which incomplete or occluded information is incrementally inferred rather than overwritten wholesale.

3.4 Sudoku-style reasoning case: resolving ambiguity

Consider an egocentric scenario where targets interact under severe occlusion:

- **t_1 (Initialization):** Target A (deep voice) and Target B (high voice) are tracked. Both walk behind a partition. Visual tokens are lost.
- **t_2 (Conflict detection):** A mixed acoustic signal arrives. TDOA features indicate a sound source at 45° (the occluded zone). The candidate pool for the source is $\{v^{(A)}, v^{(B)}\}$.
- **t_3 (Active hypothesis testing):** The LLM generates the targeted mask tokens. The external decoder isolates the audio stream.
- **t_4 (Delta-update and global consistency):** The re-injected audio reveals a "high voice" signature. The internal constraints are checked: $S^{(A)}{}_{voice}$ contradicts the observation. Thus, $v^{(A)}$ is eliminated. A delta-update (ΔV_t) is outputted: modifying $v^{(B)}$'s spatial slot to 45° (speaking), and adding a Note_{ep} to $v^{(A)}$ as "silent, location inferred via exclusion." This process corresponds to a single-step elimination in a constraint propagation loop.

4. Training paradigm: supervised structured learning and critic ranking

The training pipeline is structured in progressive phases designed to systematically build the model's capacity for complex cross-modal reasoning. Each phase introduces increasing levels of temporal and semantic complexity, ensuring that the model acquires foundational skills before tackling more demanding tasks. Crucially, by strictly isolating single-window training from long-horizon aggregation, the student model and the critic are inherently prevented from exploiting future information, thereby avoiding indirect temporal leakage. This disciplined separation of training objectives is essential for ensuring that the model generalizes correctly to unseen sequences rather than overfitting to the specific temporal patterns present in the training data.

Phase 1: Single-window supervised structured learning. The pipeline begins with strict teacher-forced alignment on isolated temporal windows. The LLM is trained to map multimodal intents to specific actions, including intent-to-mask generation, and to generate valid delta-updates (ΔV_t) based solely on the current window's context. This single-window training is essential for establishing the physical boundaries of the state encoder and ensuring that the model learns fundamental alignment tasks without relying on extended temporal context. During this phase, the model develops a precise understanding of the relationship between visual and acoustic features, learning to produce structured state representations that accurately reflect the current observational evidence. By constraining the model to a single temporal window, this phase prevents the premature development of long-horizon dependencies that could introduce noise or instability into the learning process. The teacher-forced supervision provides a strong and reliable training signal, enabling the model to quickly converge to accurate state representations.

Phase 2: Critic-guided hypothesis ranking. Once stable behavior is established in Phase 1, a lightweight multimodal transformer critic is introduced to evaluate hypothesis selection under ambiguity. This critic assesses the full generation trajectory, assigning a comprehensive score $S = C(Video, Audio, R_t, M_t, \Delta V_t)$, where R represents the explicit reasoning text and M the mask tokens. The critic evaluates four core dimensions that together capture the quality of the model's reasoning process.

- First, audio-visual mask alignment is assessed to determine whether the generated mask accurately

isolates the acoustic target matching the visual or spatial intent.

- Second, profile update rationality is scrutinized to verify that the delta-updates are logically sound given the re-injected evidence.
- Third, cross-modal physical consistency is evaluated to ensure that the proposed state matrix resolves contradictions without violating physical constraints such as spatial exclusivity.
- Fourth, explicit reasoning quality is assessed to confirm that the chain-of-thought is logical, physically grounded, and aligned with the final delta-update.

The student LLM is subsequently fine-tuned via DPO-style preference optimization or RL reward maximization to optimize this comprehensive critic score, reinforcing cross-modal consistency reasoning rather than simple pattern imitation. This phase is critical for developing the model's capacity to handle ambiguous and conflicting multimodal inputs, as it directly incentivizes the production of logically coherent and physically consistent state estimates.

Phase 3: Multi-window long-horizon supervision. Finally, the generation is unrolled across long videos to test the model's ability to maintain global state consistency over extended temporal horizons. To ensure that local delta-updates compound correctly into a globally consistent state over time, direct supervision is applied on the final aggregated constraint graph against the sequence-end ground truth. This objective minimizes the terminal state divergence, forcing the model to maintain long-term coherence and robust object permanence.

$$min\| G_T - G_T^{GT}\| \tag{3}$$

By progressively increasing the temporal scope and complexity of the reasoning tasks across the three phases, this multiphase training paradigm effectively transforms the LLM into a reliable consistency solver capable of navigating highly dynamic and partially observable environments. The terminal supervision signal ensures that the model does not merely optimize local consistency at each timestep, but rather learns to produce globally coherent state trajectories that accurately reflect the true evolution of the observed scene.

5. Conclusion

The application of LLMs in embodied perception is fundamentally redefined by APUT. By explicitly positioning the LLM as a cross-modal consistency reasoning engine rather than a mere tracker, long-term identity maintenance is framed as a solvable constraint network. This shift in perspective allows for the integration of logical reasoning into the core of the perception pipeline, addressing the limitations of passive alignment models that collapse under contradictory or absent sensory evidence. Through explicit delta-update state estimation, the model can efficiently manage and update complex state matrices without the computational burden of full regeneration at every timestep, preserving persistent context while remaining responsive to new observations. Furthermore, the active hypothesis testing mechanism empowers the system to actively resolve ambiguities by seeking clarifying information, closely mirroring the processes by which human cognition maintains object permanence in complex and dynamic environments [1]. The critic-guided supervised structured learning pipeline ensures that the model develops robust reasoning capabilities, grounded in physical commonsense and logical coherence, rather than superficial pattern matching. Ultimately, it is demonstrated that true object permanence in AI arises not from passive observation, but from internal logical global coherence, providing a robust and scalable solution for unconstrained egocentric state estimation. This paradigm offers a promising direction for future research in embodied AI, moving beyond simple tracking towards deep, reasoning-based perception that can operate reliably in the open-world conditions encountered by real-world autonomous systems.

Looking ahead, several promising avenues for future development can further enhance the APUT framework and extend its applicability. One critical direction involves scaling the model to handle an increasing number of interacting entities and more complex acoustic environments, potentially by integrating hierarchical state representations or more efficient constraint propagation algorithms. Additionally, while the current architecture relies on a deterministic external decoder for audio separation, end-to-end differentiable integration of the separation module could allow for joint optimization of both reasoning and low-level perception, potentially yielding more robust feature extraction in extremely noisy set-

tings. Another vital area for exploration is the adaptation of APUT for real-time embodied agents, such as autonomous robots or wearable assistive devices, which necessitates minimizing inference latency and optimizing the delta-update mechanism for continuous streaming data. Finally, expanding the cross-modal reasoning engine to incorporate additional sensory modalities—such as tactile or depth information—could provide a more comprehensive and resilient understanding of the physical world, moving closer to true human-like multisensory object permanence.

References

[1] Donley, J., et al. (2020). EasyCom: An augmented reality dataset to support algorithms for easy communication. *arXiv preprint arXiv:2007.03222*.

[2] Wang, Z., et al. (2026). TeleMem: Building long-term and multimodal memory for agentic AI. *arXiv preprint*.

[3] Zhao, Y., et al. (2024). VideoPrism: A foundational visual encoder for video understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Meta FAIR. (2024). Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

[5] Liu, H., et al. (2024). AudioSep: Separate anything you describe. *International Conference on Learning Representations (ICLR)*.

[6] Chen, X., et al. (2025). Slot-VLM: Object-centric visual language models for persistent tracking. *arXiv preprint*.

[7] Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2), 175-219.