CORE

Cognitive Objects Representation Engine

Stateful World Representation for Reliable Vision and language AI

Technical Report

Boltzmind.ai
Coreworldmodel.com

March 2026

Executive Summary

Modern vision and vision-language systems process video as disconnected frames. While they demonstrate impressive zero-shot capabilities, they lack persistent internal state. Objects are rediscovered each frame, semantic labels fluctuate under minor perturbations, and downstream reasoning systems operate without a coherent temporal model of the world.

CORE (Cognitive Objects Representation Engine) introduces a World State Layer for AI systems. CORE maintains a dynamic pool of persistent Object Kernels, latent representations that evolve over time and encode object identity, state, and relationships.

Benchmark Highlights: 66.1% MOTA on MOT17 (structural tracking), Semantic Stability improved from 86.7% to 100% under perturbation, and accumulation of 3,800+ temporal relationships.

1. The Stateless Perception Problem

State-of-the-art Vision-Language Models analyze video frame-by-frame without maintaining persistent internal state. This results in loss of object continuity, semantic flicker, redundant recomputation, and lack of structured temporal grounding.

2. Architecture: The Kernel Pool

CORE maintains a dynamic collection of Object Kernels. Each kernel evolves according to $h_t = \mathrm{GRU}(z_t, h_{t-1})$, ensuring temporal continuity. A self-supervised consistency loss enforces identity inertia and prevents drift.

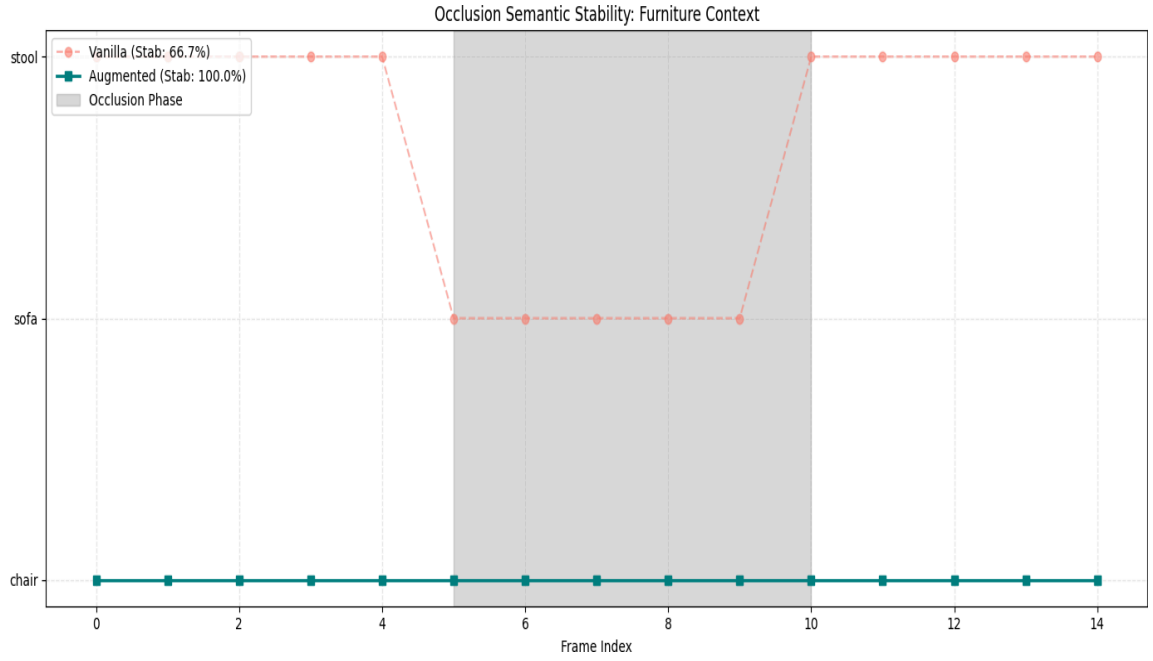
3. Evaluation I: Structural Tracking (MOT17)

Metric	Score	Interpretation
MOTA	66.1%	Strong spatial continuity and object presence enforcement
IDF1	2.9%	Limited long-term re-identification

4. Evaluation II: Semantic Stability Under Perturbation

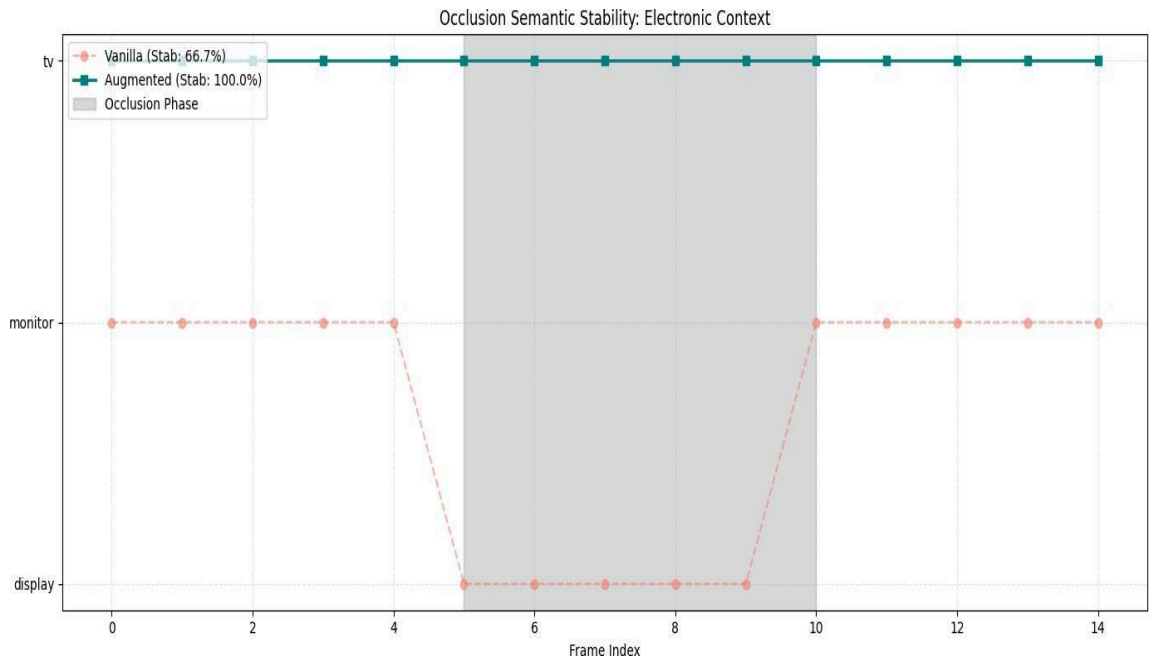
System	Semantic Stability
Vanilla CLIP	86.7%
CORE-Augmented	100%

Electronic Context



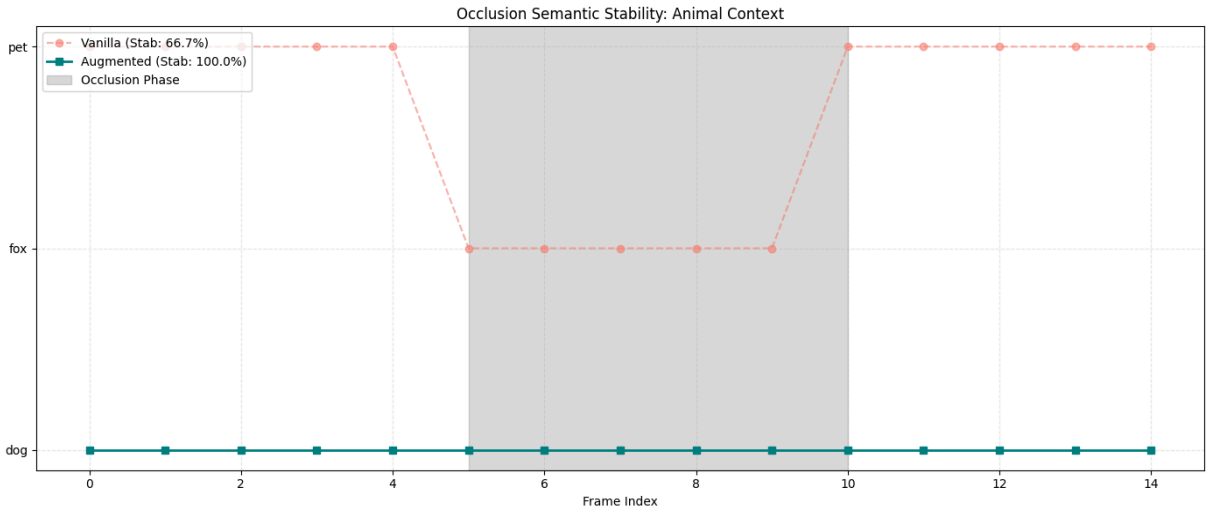
Observation: Vanilla recognition models collapse to incorrect semantic states during occlusion (66.7% stability), whereas CORE maintains 100% semantic stability.

Furniture Context



Observation: CORE preserves object identity across occlusion phases with complete stability, demonstrating context-independent persistence.

Animal Context



Observation: CORE maintains object identity across 5 frames of perturbation.

5. Relational Accumulation

CORE enables persistent entity graphs and temporal state transitions. In benchmark runs, the system accumulated over 3,800 structured relationships, forming a dense and queryable world model.

6. Strategic Implication: World State as Infrastructure

CORE transforms raw video streams into structured temporal JSON representing persistent entities and relations. It enables queryable object history, event-driven reasoning, and grounded language interaction. CORE acts as stateful RAM for perception, providing foundational infrastructure for embodied AI systems.

7. Implications for Robotics and Embodied AI

Persistent semantic identity is foundational for robotic manipulation, planning, and reasoning. CORE enables memory-grounded perception required for real-world autonomy.

Applications include:

- Autonomous manipulation under visual disruption
- Stateful world modeling for embodied agents
- Long-horizon task execution
- Human-robot interaction requiring object continuity