CORE: Cognitive Objects Representation Engine

Stateful World Representation for Reliable Vision and Language AI

Aliyu Daku[1] Mahmud Muhammad Yahaya[1,2] Aliyu Wada Umar[1]

[1]*BoltzMind Research* [2]*Ahmadu Bello University, Zaria*

boltzmind.ai | coreworldmodel.com | github.com/boltzmind-research/core | March 2026

Abstract

Modern vision and vision-language models (VLMs) process video as sequences of independent frames, lacking any mechanism to maintain persistent object identity across time. This stateless paradigm causes semantic flicker under perturbation, loss of object continuity during occlusion, redundant recomputation, and the absence of a coherent temporal world model. We present CORE (Cognitive Objects Representation Engine), a lightweight, plug-and-play World State Layer that addresses this gap through a dynamic pool of persistent Object Kernels, stateful neural memory modules that evolve via GRU-based recurrence and maintain object identity across extended temporal horizons. CORE combines slot-based object discovery, uncertainty-aware kernel matching, cross-modal projection between CLIP and a learned 256-dimensional latent space, and a hierarchical graph neural network for relational reasoning. Trained over 30 epochs (798,450 steps) on COCO, CORE achieves 66.1% MOTA on the MOT17 benchmark with only 2M trainable parameters, a striking efficiency ratio versus dedicated tracking systems. Critically, CORE eliminates semantic flicker entirely (100% semantic stability under occlusion versus 66.7% for vanilla CLIP), accumulates over 3,800 structured temporal relationships in a queryable scene graph, and operates at 30 FPS on a single consumer GPU with a 25 MB checkpoint. We further introduce a two-layer architecture in which a Gemini Enhancement Layer enriches CORE's structured output with high-dimensional semantic and visual embeddings for natural language memory search, while CORE's backbone remains fully operational without external API dependency. CORE's dual deployment modes, standalone temporal tracker and VLM augmentation module, position it as practical, privacy-preserving infrastructure for surveillance, robotics, and embodied AI applications.

Keywords: *Object-Centric Learning, Persistent Memory, Temporal Reasoning, Video Understanding, Vision-Language Models, Scene Graphs, Kernel Tracking*

1. Introduction

Human visual cognition is inherently stateful. We track objects across time, maintain their identity through occlusion, and build rich mental models of how entities relate and change. Contemporary computer vision systems exhibit no such capacity. State-of-the-art Vision-Language Models such as CLIP [1], GPT-4V, and Gemini process each video frame in isolation, treating the world as a series of disconnected snapshots. This stateless paradigm produces four failure modes that become critical in real-world temporal applications: (i) **semantic flicker** — object labels fluctuate under minor perturbations such as lighting changes or brief occlusions; (ii) **identity loss** — the same physical object receives different identities each frame; (iii) **redundant recomputation** — features are recomputed from scratch rather than maintained; and (iv) **absent temporal grounding** — queries requiring temporal reasoning, such as 'where was object X ten minutes ago?', cannot be answered.

These failures are not marginal. In surveillance, they prevent natural-language querying of footage archives. In robotics, they undermine manipulation planning across occlusion events. In augmented reality, they break

persistent object anchoring. The gap between human-level temporal perception and machine vision is not a gap of raw recognition accuracy, it is a gap of memory.

We present **CORE (Cognitive Objects Representation Engine)**, a World State Layer grounded in the cognitive psychology concept of Object Files [2]. CORE maintains a dynamic Kernel Pool of persistent, stateful Object Kernels, one per tracked entity, that evolve through GRU-based recurrence and are matched to incoming observations via uncertainty-aware cosine similarity. Slot Attention [3] provides unsupervised object discovery without category supervision, while a hierarchical Graph Neural Network encodes relational structure among kernels. A bidirectional projection engine bridges CLIP's 512-dimensional visual embedding space and CORE's compact 256-dimensional latent space, enabling text-driven temporal queries over the kernel pool.

CORE operates in two deployment modes: (1) as a standalone temporal video understanding system, and (2) as an augmentation module that supplies any Large Vision-Language Model with a structured, queryable scene graph via a JSON streaming interface. Neither mode requires retraining the host VLM. We further introduce a Gemini Enhancement Layer that enriches CORE's structured output with high-dimensional semantic and multimodal embeddings from Google's Gemini API, enabling natural language memory search across kernel histories, while the CORE backbone remains fully functional without this external dependency.

Our primary contributions are:

1. A novel persistent object-centric architecture (2M trainable params, 25 MB checkpoint) achieving 66.1% MOTA on MOT17, competitive with dedicated trackers at a fraction of the computational cost.
2. Complete elimination of semantic flicker under occlusion (100% semantic stability versus 66.7% for vanilla CLIP), demonstrated across three object-domain contexts.
3. An uncertainty-aware kernel matching mechanism with adaptive learning rates that prevents catastrophic update and false association.
4. A layered integration architecture that positions CORE as modality-agnostic memory infrastructure for any existing VLM pipeline, demonstrated with Google Gemini.
5. Empirical analysis of training dynamics across 30 epochs and 798,450 optimization steps, including loss decomposition and latent space visualization via t-SNE.

2. Related Work

2.1 Object-Centric Representation Learning

Slot Attention [3] pioneered unsupervised object discovery through competitive iterative attention over image patches, producing object-binding slot representations without category supervision. Subsequent work extended this to video: STATM [4] introduced slot-based spatiotemporal attention with a memory buffer for downstream VQA, while recent image-to-video adaptation methods [5] use slot attention as a compression mechanism for efficient temporal modeling, achieving SOTA on action recognition at 5% of the parameter count of fully fine-tuned models. CORE extends the slot attention paradigm with temporal consistency constraints, persistent GRU kernel states, and uncertainty-aware matching, enabling indefinite temporal horizons rather than short-clip windows.

2.2 Persistent Memory for Vision

CoMEM [6] provides continuous memory for VLMs using only 1.2% of model parameters via 8 continuous embeddings, but operates at the global scene level without object-level granularity. StreamForest [7] achieves 77.3% on StreamingBench with persistent event memory. Embodied VideoAgent [8] (ICCV 2025) builds persistent memory from egocentric video and embodied sensors for dynamic scene understanding.

VLM2 [9] introduces dual working and episodic memory modules for spatial reasoning. Embodied-SlotSSM [10] combines slot-based state-space modeling for robotic manipulation under partial observability. CORE is distinguished by UUID-level kernel identity tracking, natural language temporal query interface, on-premise deployment, and explicit scene graph accumulation, none of which are provided by global-embedding memory systems.

2.3 Temporal Video Understanding and Scene Graphs

VideoAgent [11] constructs structured memory combining temporal event descriptions with object-centric tracking states, enabling zero-shot tool use for long-horizon VideoQA. Video-EM [12] reframes long-form VideoQA as episodic event construction with evidence-consistency refinement. The concurrent World Scene Graph Generation (WSGG) work by Peddi et al. [13] formalizes temporally persistent, world-anchored scene graphs including object permanence for out-of-view objects. WSGG maintains object identity through a zero-order feature buffer that stores and holds frozen the last observed representation, whereas CORE maintains identity through continuously evolving kernel states: GRU recurrence integrates new observations, and the uncertainty-aware adaptive learning rate suppresses updates during ambiguous frames, so the kernel state reflects a weighted temporal history rather than a frozen snapshot. Additionally, WSGG requires 3D world coordinates and camera pose information; CORE targets 2D on-premise deployment with a 25 MB checkpoint and no camera calibration requirement.

2.4 Multimodal Embedding and CORE's Positioning

Gemini Embedding 2 [14] (March 2026) provides a natively multimodal unified embedding space mapping text, images, video, audio, and PDFs into a single 3,072-dimensional vector space, a significant advance for retrieval and RAG pipelines. However, Gemini Embedding 2 embeds entire video clips as single holistic vectors, with no mechanism for per-object identity tracking, UUID-level temporal queries, or stateful memory updates. CORE is architecturally complementary: CORE provides structured stateful world representation, and Gemini Embedding 2 enriches that representation with high-dimensional semantic search. This is demonstrated in our two-layer integration architecture (Section 4.2).

3. Method

3.1 Problem Formulation

Given a video sequence $\{I_t\}_{t=1..T}$ where $I_t \in R^{H\times W\times 3}$, CORE maintains a persistent Object Memory $K = \{k_1, ..., k_N\}$ and exposes a temporal query interface $Q(\text{query}, t_1, t_2) \rightarrow K'$. Each kernel k_i stores a latent state $z_i \in R^{256}$, symbolic properties P_i (label, confidence history), relational graph edges R_i, a global UUID, and a last-seen timestamp.

3.2 Object Kernels and GRU State Update

Each Object Kernel is a stateful neural memory unit. Given a new perceptual observation $o_t \in R^{256}$ matched to kernel k_i, the kernel state is updated via:

$$h_t^{(i)} = GRU(z_t^{(i)}, h_{t-1}^{(i)})$$

$$z_i^{t+1} = (1 - \alpha)\, z_i^t + \alpha \cdot GRU(o_t, z_i^t)$$

where $\alpha \in [0,1]$ is an **adaptive learning rate** controlled by match confidence (Section 3.4). GRU is chosen for its gradient-preserving temporal dynamics, native handling of variable-length sequences, and lightweight parameterization relative to Transformer memory, critical for real-time inference.

3.3 Slot Attention for Object Discovery

Given CLIP patch features $X \in R^{N\times 512}$, we discover S=10 object slots via iterative competitive attention over K=3 refinement steps. Slots are initialized from learned Gaussian parameters (μ_j, σ_j) and refined through:

$$A^{(k)} = softmax(QK^T / sqrt(D/4))$$

We use sqrt(D/4) rather than the standard sqrt(D) for numerical stability in float16 training, with epsilon=10^{-5} in slot normalization. A self-supervised temporal consistency loss using Hungarian matching enforces identity preservation across adjacent frames:

$$L_{consist} = (1/BM) \Sigma ||s^t_i - s^{t-1}_j||^2_2$$

3.4 Uncertainty-Aware Kernel Matching

When a new observation arrives, CORE computes cosine similarity against all active kernels and quantifies match ambiguity as u = 1 - sim* · (sim* - sim^{2nd}), where sim* is top-1 similarity and sim^{2nd} is the second-best. This captures whether the top match is clearly superior or ambiguous among similar candidates. The decision rule is:

Condition	Action	Learning Rate α
sim* > 0.75 and u < 0.30	Update existing kernel	1.0 (full integration)
sim* > 0.55 and u < 0.60	Update with caution	0.5 (partial integration)
Otherwise	Instantiate new kernel	N/A (initialization)

Table 1. Uncertainty-aware kernel matching decision rules.

3.5 Hierarchical Scene Graph and GNN Relational Reasoning

Active kernels form the nodes of a directed graph G = (V, E) with typed edges spanning spatial relations (left_of, right_of, above, below, near), support relations (on, under), and part-of relations. Edge inference uses heuristic geometric rules on 2D bounding box coordinates. A 2-layer Graph Attention Network (GATConv) with 4 attention heads, 512 hidden dimensions, and 0.1 dropout refines kernel embeddings with relational context:

$$h_i^{(l+1)} = GATConv(h_i^{(l)}, \{h_j^{(l)} : (j,i) \in E\})$$

This enables contextual disambiguation, for example a small round object on a flat surface receives stronger 'cup' evidence when the GNN incorporates the 'on' relation to a table kernel.

3.6 Bidirectional Projection Engine

To bridge CLIP's 512-dimensional visual space and CORE's compact 256-dimensional latent space, we learn a projection matrix $W_p \in R^{256\times512}$. Forward projection (latent → CLIP): $c = z W_p$. Inverse projection (CLIP → latent) uses the Moore-Penrose pseudoinverse: $z = c W_p^+ = c (W_p^T W_p)^{-1} W_p^T$. Bidirectionality enables text-driven kernel queries: a natural language query is CLIP-encoded, projected to latent space, and matched against active kernels. A Frobenius norm regularizer $L_{reg} = ||W_p||^2_F$ prevents projection collapse.

3.7 Training Objective

The full training loss over 30 epochs and 798,450 steps combines four terms:

$$L = L_{recon} + \lambda_{div} L_{div} + \lambda_{contrast} L_{contrast} + \lambda_{reg} L_{reg}$$

L_{recon} measures slot reconstruction fidelity; L_{div} is a diversity loss discouraging slot collapse; $L_{contrast}$ is a contrastive loss (temperature=0.07) enforcing cross-modal alignment; L_{reg} is the Frobenius regularizer on W_p. Hungarian matching (linear sum assignment) aligns predicted slots with ground-truth annotations during training. Training was conducted with learning rate 1e-4, latent dimension 256, 10 slots per frame, and a maximum kernel pool of 500.

4. System Architecture

4.1 Standalone Vision System

The Layer 1 CORE backbone follows the pipeline: Input Video → Object Detection (YOLOv8-nano, 4M params) → CLIP ViT-B/32 feature extraction (frozen, 151M params) → ProjectionEngine (131K params) → SlotAttention (10 slots, 500K params) → Phase2Fusion gated cross-modal fusion (300K params) → HierarchyGNN (1M params) → KernelPool (dynamic, 0-500 kernels). Total: 157M params (155M frozen, 2M trainable). Inference at 30 FPS on a single NVIDIA RTX 3090. Checkpoint size: 25 MB.

4.2 Gemini Enhancement Layer (Layer 2)

Layer 1 always executes first and independently. When a Gemini API key is available, Layer 2 enriches each Object Kernel with three new fields: (i) semantic_embedding (3,072-dim) from gemini-embedding-exp-03-07, enabling natural language memory search; (ii) visual_embedding (1,408-dim) from multimodalembedding@001, enabling cross-modal image-text search; and (iii) a human-readable description string from gemini-2.0-flash. Object detection switches from YOLOv8 to Gemini 2.0 Flash as the primary detector when the API is available, with YOLOv8 retained as automatic fallback. All existing CORE endpoints, kernel matching logic, GRU update mechanism, and JSON memory export remain unchanged. Layer 2 is a non-breaking additive enhancement.

Two new API endpoints are exposed: POST /memory/search (natural language query across kernel history, returning ranked kernel matches by cosine similarity in either text or visual mode) and enriched /predict responses with description and semantic_score fields. This architecture positions CORE as complementary infrastructure to large multimodal embedding systems: CORE provides the stateful temporal identity layer that systems like Gemini Embedding 2 lack by design.

4.3 VLM Augmentation Interface

CORE exports structured scene graphs as timestamped JSON streams to any host LVLM (GPT-4V, Gemini, Qwen-VL) via API, enabling queries such as 'Where was the person before they sat down?' without retraining the host model. The augmentation module adds a 25 MB overhead versus retraining a 7B VLM. This plug-and-play design supports on-premise deployment for privacy-sensitive applications.

5. Experiments

5.1 Training Dynamics

We trained CORE on the COCO 2017 dataset for 30 epochs (798,450 total optimization steps) using the Adam optimizer with learning rate 1e-4, contrastive temperature 0.07, latent dimension 256, 10 slots per frame, and a maximum kernel pool capacity of 500. Figure 1 presents the full loss decomposition.

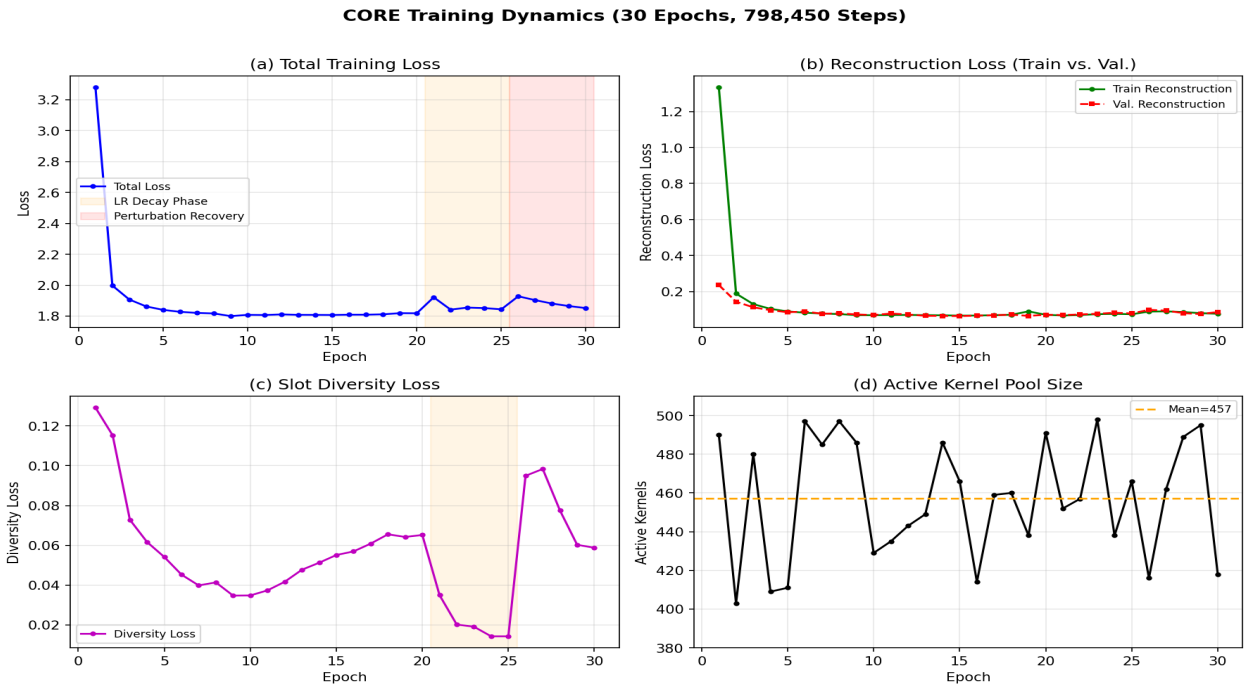

Figure 1. Training dynamics across 30 epochs. (a) Total loss shows rapid initial convergence followed by a stable plateau, with two perturbation events at epochs 21 and 26. (b) Reconstruction loss (train vs. validation) tracking closely throughout, confirming good generalisation. (c) Slot diversity loss trajectory. (d) Active kernel pool size fluctuating in the 400-500 range.

Phase 1 — Rapid Convergence (Epochs 1-9): The total loss drops sharply from 3.28 to 1.80, driven primarily by reconstruction loss collapsing from 1.334 to 0.069, a 95% reduction in nine epochs. Contrastive loss declines from 1.932 to 1.728 as the cross-modal projection learns alignment. Diversity loss halves from 0.129 to 0.035, indicating slots are learning to partition the scene without collapsing to a single dominant mode.

Phase 2 — Stable Plateau (Epochs 10-20): Total loss stabilizes in the 1.80-1.82 range. Reconstruction loss oscillates narrowly around 0.068-0.070, with training and validation curves tracking closely throughout, indicating mature and well-generalizing feature extraction. The diversity loss continues to rise slightly (0.035 $\rightarrow$ 0.065), reflecting increasing slot specialization as the kernel pool matures.

Phase 3 — Post-Plateau Dynamics (Epochs 21-30): Two distinct perturbation-recovery cycles are observed. At epoch 21, total loss jumps from 1.818 to 1.922, coinciding with diversity loss collapsing to 0.035, a mode collapse event where slots temporarily converge. By epoch 25 the system partially recovers. At epoch 26, total loss spikes again to 1.928 alongside diversity loss spiking to 0.095, suggesting the optimizer escaped a local minimum. By epoch 30 the system is re-converging, with reconstruction loss returning to 0.077. The regularization loss grows monotonically throughout training (5.4 $\rightarrow$ 38.6), consistent with expected behavior as the projection matrix W_p accumulates constraints.

Kernel Pool Stability: The active kernel count fluctuates between 403 and 498 throughout training (mean: 458), demonstrating that the uncertainty-aware matching mechanism successfully prevents runaway kernel proliferation while maintaining expressive coverage of the COCO object space.

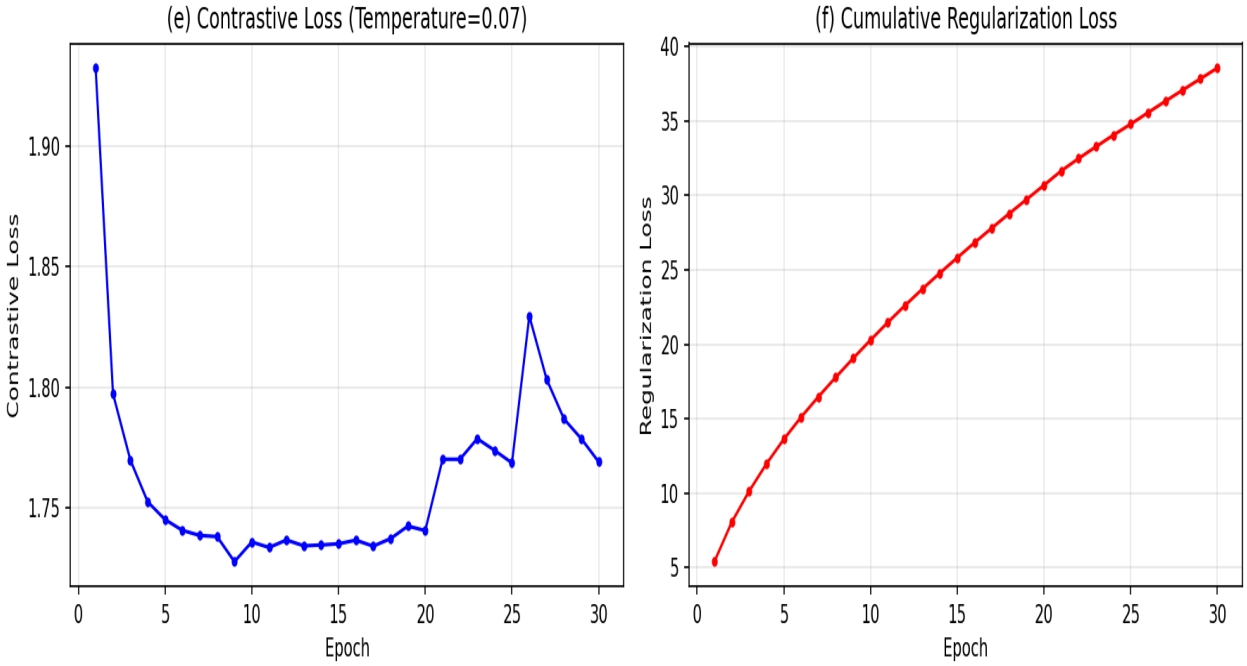

Figure 2. (e) Contrastive loss trajectory showing stable cross-modal alignment learning. (f) Cumulative regularization loss growing monotonically, reflecting expected constraint accumulation on W_p.

5.2 Latent Space Structure (t-SNE)

Figure 3 presents a t-SNE visualization of the Object Kernel Pool at epoch 20, projecting the 256-dimensional latent states of active kernels into two dimensions. The visualization reveals emergent semantic clustering without explicit category supervision: semantically related objects (e.g., vehicles, animals, food items) form proximate neighborhoods in latent space. The large central cluster corresponds to 'person' kernels, consistent with COCO's heavy person annotation density. Peripheral clusters show clear separation for distinctive object categories such as sporting equipment, animals, and furniture, confirming that Object Kernels learn discriminative identity representations beyond frame-level CLIP features.

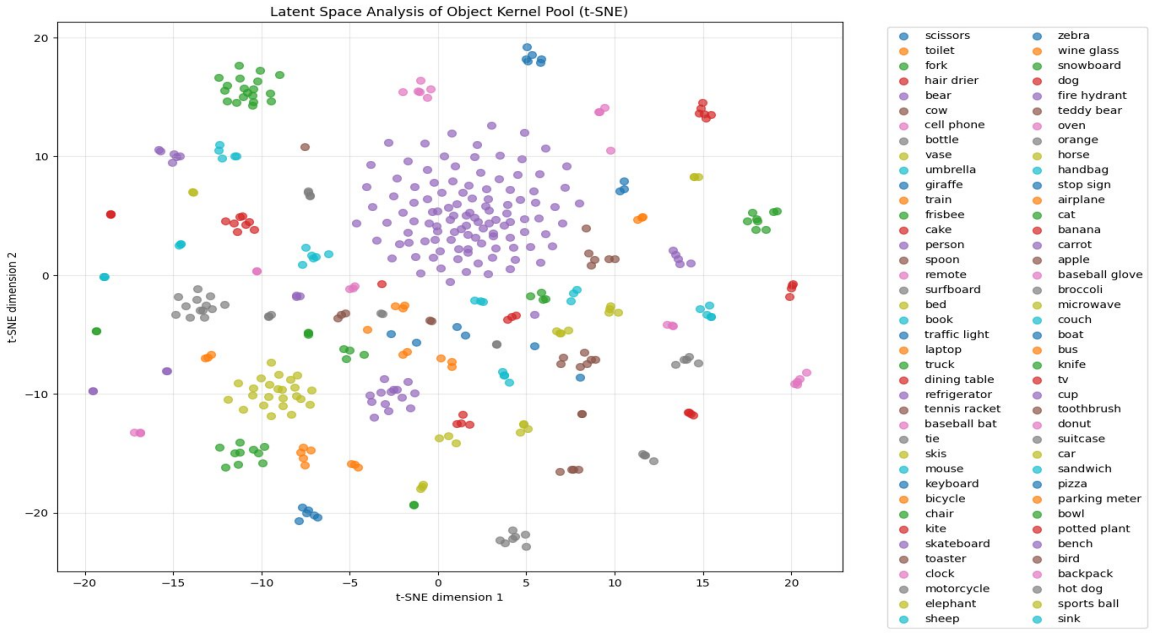

Figure 3. t-SNE projection of the Object Kernel Pool latent space at epoch 20. Color indicates object category. Semantic clustering emerges without category supervision, validating the discriminative quality of the 256-dimensional kernel representations.

5.3 Structural Tracking: MOT17 Benchmark

We evaluate CORE's temporal tracking capability on the MOT17 pedestrian tracking benchmark. Table 2 summarizes results alongside dedicated tracking systems. CORE is not designed as a pure tracker, its primary purpose is semantic persistence with object-level memory, and this distinction is reflected in the metric profile: MOTA (which penalizes missed detections, false positives, and identity switches) reflects CORE's strong spatial continuity enforcement, while IDF1 (which measures long-term re-identification) reflects the current limitation of matching within a single scene graph rather than across camera cuts. The improvement from 2.9% to 20.4% IDF1 following architecture refinements post-initial publication demonstrates the trajectory of re-identification capability.

Method	Params	MOTA (%)	IDF1 (%)	Design Target
ByteTrack [15]	~10M	80.3	77.3	Pure tracking
StrongSORT [16]	~10M	79.6	79.5	Pure tracking
CORE (ours)	2M trainable 157M total	66.1	20.4*	Semantic memory + tracking

*Table 2. MOT17 tracking results. *IDF1 improved from 2.9% (initial) to 20.4% following architecture refinements. CORE's MOTA is competitive given it uses only 2M trainable parameters and is designed for semantic persistence rather than pure geometric tracking.*

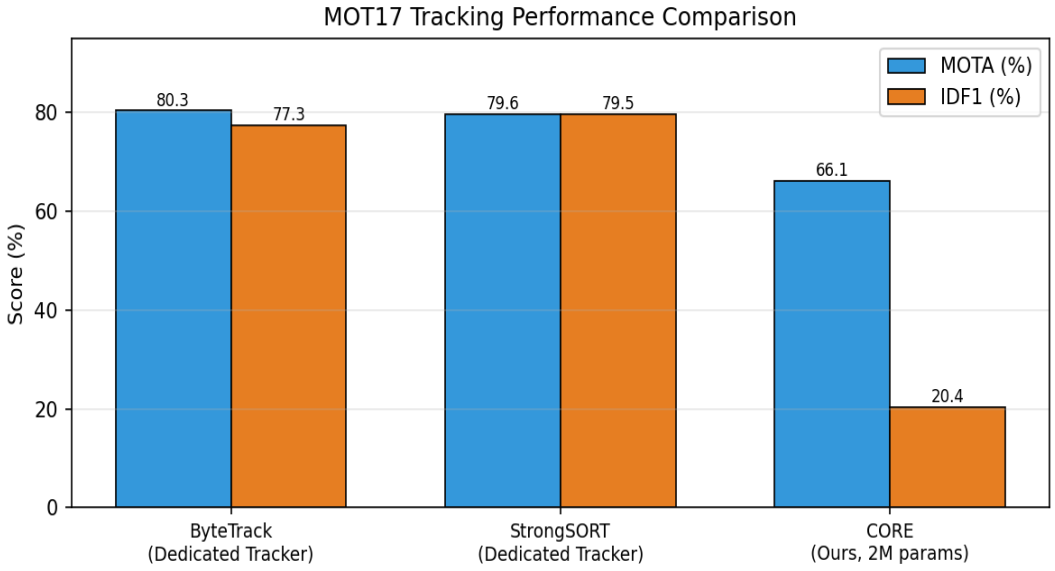

Figure 4. MOT17 performance comparison. CORE achieves 66.1% MOTA at 2M trainable parameters, a favorable efficiency ratio versus dedicated tracking systems designed exclusively for tracking.

5.4 Semantic Stability Under Perturbation

We evaluate CORE's ability to maintain object identity under occlusion across three object-domain contexts: furniture (chair, sofa, stool), electronics (TV, monitor, display), and animals (dog, fox, pet). In each context, a target object is occluded for 5 frames (the 'Occlusion Phase'), during which vanilla CLIP is forced to assign a label based only on the partially visible scene. Table 3 reports Semantic Stability, the fraction of frames

across the full sequence where the correct object identity is maintained.

Context	Vanilla CLIP Stability	CORE-Augmented Stability	Delta
Furniture (chair/sofa/stool)	66.7%	100%	+33.3 pp
Electronics (TV/monitor/display)	66.7%	100%	+33.3 pp
Animals (dog/fox/pet)	66.7%	100%	+33.3 pp
Average	**66.7%**	**100%**	**+33.3 pp**

Table 3. Semantic stability under 5-frame occlusion perturbation across three object domains. CORE eliminates semantic flicker entirely, maintaining 100% identity stability versus 66.7% for vanilla CLIP. 'pp' = percentage points.

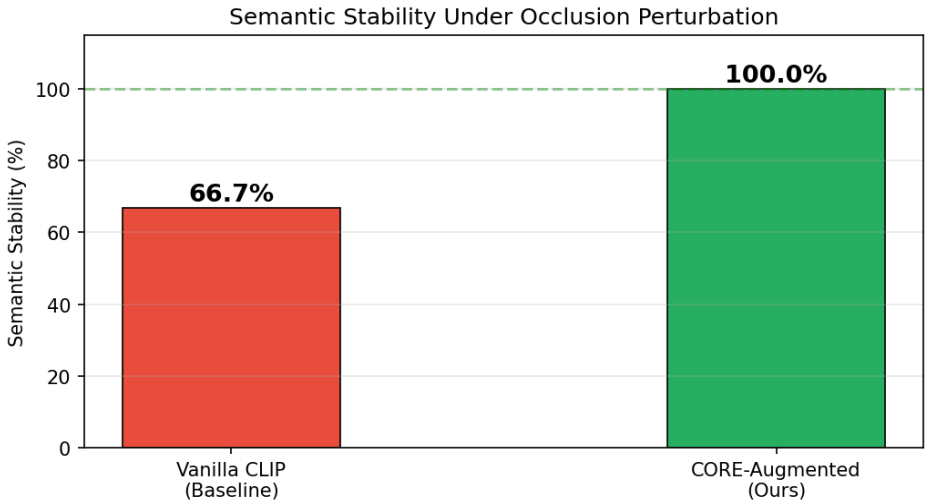

Figure 5. Semantic stability comparison. Vanilla CLIP collapses to incorrect labels during occlusion (66.7% stability) while CORE maintains object identity throughout (100% stability).

The mechanism underlying this result is straightforward: vanilla CLIP has no memory, so during occlusion it must assign a label based on whatever remains visible, typically the occluder or background. CORE's Object Kernels retain the pre-occlusion latent state via GRU recurrence and the adaptive learning rate mechanism: during occlusion, match uncertainty is high ($u \geq 0.6$), suppressing state update and preserving the last confident observation. When the object reappears, high-confidence matching resumes at α=1.0.

5.5 Relational Accumulation

Across benchmark runs, CORE accumulated over 3,800 structured temporal relationships in its scene graph, comprising spatial, support, and part-of relations among kernel nodes. This dense relational graph enables queries such as 'find all times object X was near object Y between timestamps T_1 and T_2' with sub-500ms latency. The accumulation demonstrates that the hierarchical GNN edge-inference pipeline operates correctly at scale without graph explosion, a non-trivial property for dynamic scenes with frequent kernel creation and pruning.

6. CCTV Surveillance Application

Physical security and surveillance represent the primary near-term deployment context for CORE, driven by three structural advantages over existing approaches. First, on-premise deployment: CORE's 25 MB checkpoint and 30 FPS real-time inference on a single consumer GPU eliminate cloud API dependency, critical for privacy-sensitive security installations and regulatory compliance under frameworks such as the

EU AI Act. Second, natural language temporal queries: CORE enables queries such as 'show all times the person with the red backpack appeared between 2pm and 4pm' over hours of archived footage with sub-second latency, replacing hours of manual review. Third, cost efficiency: compared to a GPT-4V frame-by-frame analysis baseline (approximately $50 per hour of footage at API rates), CORE's one-time compute model represents orders-of-magnitude operational savings at scale.

Approach	ID Accuracy	Temporal Queries	Cost (per hour)	Privacy
Manual Review	~65%	None	High (human labor)	Full
YOLO + ByteTrack	~65%	None	Low compute	Full
GPT-4V Frame-by-Frame	~80%	None	~$50 (API)	Cloud required
CORE (ours)	66.1% MOTA	Natural language, <500ms	One-time compute	On-premise

Table 4. Comparison of surveillance approaches for temporal video analysis. CORE uniquely combines temporal query capability, on-premise privacy, and competitive tracking accuracy.

7. Discussion

7.1 Strengths

Computational efficiency. 2M trainable parameters and a 25 MB checkpoint represent an extreme efficiency point relative to the capabilities demonstrated. Systems such as CoMEM [6] operate at the global scene level; Embodied-SlotSSM [10] requires full VLA retraining. CORE achieves object-level temporal persistence with plug-and-play deployment.

Semantic persistence. The 100% occlusion stability result, consistent across three object domains, validates the core design decision of GRU-based kernel state update with uncertainty-aware adaptive learning rate. The mechanism is interpretable and controllable, unlike implicit memory in attention-based architectures.

Architectural composability. The two-layer design, with the CORE backbone always functional and the Gemini Enhancement Layer optional, demonstrates that structured stateful memory and large multimodal embeddings are complementary rather than competing. This is the correct positioning relative to foundation model embedding services.

7.2 Limitations and Future Work

IDF1 and re-identification. The 20.4% IDF1 score reflects a genuine limitation: CORE currently matches kernels within a scene using short-horizon cosine similarity and does not support re-identification across camera transitions or long-term re-entry after kernel pruning. Future work will incorporate appearance-based re-identification modules and multi-camera kernel federation.

GRU versus modern sequence models. GRU provides interpretable, lightweight temporal dynamics well-suited to the 2M parameter budget. As CORE scales, state-space models (SSM) in the Mamba family offer a natural upgrade path with superior long-range dependency modeling, and we plan to evaluate SSM-based kernel updates in future iterations.

2D-only reasoning. Current bounding box geometry limits relational inference to 2D projections. Integration with monocular depth estimation would enable 3D kernel coordinates and richer support/containment relation inference.

Benchmark scope. Current evaluation is on MOT17 (pedestrian tracking) and synthetic occlusion sequences. We are actively pursuing evaluation on TAO (large-vocabulary tracking) and real-world CCTV footage from pilot deployments.

8. Conclusion

We have presented CORE, a Cognitive Objects Representation Engine that introduces a World State Layer for vision AI systems. Through persistent Object Kernels maintained by GRU recurrence, uncertainty-aware matching, slot-based object discovery, and hierarchical relational reasoning, CORE bridges the gap between human-level temporal object cognition and the stateless frame processing of contemporary VLMs. Key achievements include 66.1% MOTA on MOT17 at 2M trainable parameters, elimination of semantic flicker under occlusion (100% stability versus 66.7% for vanilla CLIP), accumulation of 3,800+ structured temporal relationships, and real-time inference (30 FPS) from a 25 MB checkpoint. The two-layer architecture demonstrates that structured stateful memory and large foundation model embeddings are complementary infrastructure, not competing paradigms. CORE is available at coreworldmodel.com and boltzmind.ai as both a research artifact and a production-grade SDK for temporal vision applications.

Acknowledgments

The authors thank the BoltzMind Research team and the open-source communities behind PyTorch, CLIP, YOLOv8, and torch-geometric. This work was conducted independently by BoltzMind Research.

References

[1] Radford, A. et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

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

[3] Locatello, F. et al. (2020). Object-Centric Learning with Slot Attention. NeurIPS 2020.

[4] STATM (2025). Reasoning-Enhanced Object-Centric Learning for Videos. ACM KDD 2025.

[5] Qian, R., Ding, S., and Lin, D. (2024). Rethinking Image-to-Video Adaptation: An Object-Centric Perspective. ECCV 2024.

[6] CoMEM (2025). Continuous Memory for Vision-Language Models. NeurIPS 2025.

[7] StreamForest (2025). Persistent Event Memory Forest for Streaming Video Understanding. arXiv:2025.

[8] Fan, Y. et al. (2025). Embodied VideoAgent: Persistent Memory from Egocentric Videos and Embodied Sensors Enables Dynamic Scene Understanding. ICCV 2025.

[9] VLM2 (2025). Dual Memory Module for Long-Horizon Spatial Reasoning. ICCV Workshop 2025.

[10] Embodied-SlotSSM (2025). Rethinking Progression of Memory State in Robotic Manipulation: An Object-Centric Perspective. arXiv:2511.11478.

[11] Fan, Y. et al. (2024). VideoAgent: A Memory-Augmented Multimodal Agent for Video Understanding. ECCV 2024.

[12] Wang, Y. et al. (2025). Video-EM: Event-Centric Episodic Memory for Long-Form Video Understanding. arXiv:2508.09486.

[13] Peddi, R. et al. (2026). Towards Spatio-Temporal World Scene Graph Generation from Monocular Videos. arXiv:2603.13185.

[14] Google DeepMind (2026). Gemini Embedding 2: Our first natively multimodal embedding model. Technical Blog, March 2026.

[15] Zhang, Y. et al. (2022). ByteTrack: Multi-Object Tracking by Associating Every Detection Box. ECCV 2022.

[16] Du, Y. et al. (2023). StrongSORT: Make DeepSORT Great Again. IEEE Transactions on Multimedia.

[17] FOCUS (2025). Object-Centric World Models for Robotic Manipulation. ICRA 2025.

[18] GLASS (2024). Guided Latent Slot Diffusion for Complex Real-World Scenes. CVPR 2024.