

# The Embodied Reasoning Nexus: Integrating Raw Multi-Modal Sensor Fusion with the Hierarchical Reasoning Model (HRM)

---

## I. Strategic Rationale: The Necessity of Embodied Input for Efficient Sapience

The pursuit of artificial sapience necessitates a departure from purely symbolic or language-centric architectures towards systems capable of processing and engaging with the physical environment in a causal, experienced manner. The integration of high-fidelity, raw, multi-modal sensor data, rather than relying on abstract, semantically filtered human inputs, represents a crucial paradigm shift required to anchor abstract reasoning within environmental phenomenology.

### 1.1. Reframing the Input Problem: From Semantic Inputs to Environmental Phenomenology

Traditional large language models (LLMs) operate primarily on data passively collected from the internet, which lacks the inherent causal and spatio-temporal coherence of real-world experience.<sup>1</sup> The proposed move to dedicated sensor inputs—specifically camera, microphone, accelerometer, gyroscope (IMU), and GPS—fundamentally repositions the artificial intelligence agent from a passive observer to an entity capable of **Active Agency**.<sup>1</sup> Goal achievement, particularly in complex, sequential reasoning tasks that the Hierarchical Reasoning Model (HRM) is designed to solve<sup>2</sup>, relies fundamentally on establishing a causal understanding between an agent's actions and the resulting sensory consequences.<sup>1</sup> Without inputs derived from experience, the model cannot develop the predictive capabilities necessary for long-term goal optimization.<sup>3</sup>

This transition requires the initial input layer to handle information at a raw, non-semantic level. Research in biological plausibility, for instance, confirms that the encoding of linguistic information in the brain progresses hierarchically, starting from low-level representations like

the *spectrogram-level* for auditory input, rather than high-level word tokens.<sup>4</sup> Applying this principle to the entire multi-modal sensor suite ensures that the architecture is capable of building environmental representations incrementally, mirroring cognitive processes. Furthermore, systems built upon non-semantic fusion have already demonstrated practical efficacy, such as automatically identifying key contextual segments within lecture archives using raw IMU, camera, and microphone data, providing valuable, quantifiable feedback.<sup>5</sup> The shift to dedicated sensor data is therefore not merely an input method; it is a **precondition** for the HRM to realize its full sapient potential. The internal architecture of the HRM is designed for devising and executing complex goal-oriented action sequences.<sup>2</sup> For this planning capability to translate into effective action, the reasoning engine must be grounded in an accurate, dynamically updated model of reality. By supplying raw, experienced sensory streams, the system moves beyond static reasoning into an active, predictive world model, thereby ensuring the causal link between action, perception, and goal outcome is maintained.

## 1.2. Architectural Selection: Leveraging HRM's Efficiency and Cognitive Constraints

The proposed system mandates the utilization of the Hierarchical Reasoning Model (HRM) developed by Sapient AI as the central cognitive engine.<sup>6</sup> This core architectural decision imposes a severe constraint: the integrity of HRM's efficiency must be preserved. The HRM is exceptionally lightweight, featuring only 27 million parameters, which allows it to run effectively on standard CPUs with a minimal memory footprint (under 200MB of RAM).<sup>6</sup> This efficiency advantage differentiates it from much larger models currently used in the industry. The constraint that the perception layer must not negate the computational efficiency of the 27M-parameter core dictates a critical research focus. If the front-end sensor fusion architecture is computationally heavier than the HRM itself—demanding extensive memory, massive pre-training data, or specialized, power-intensive hardware, as is common with large vision transformers—the fundamental advantage of the lightweight core is lost. Consequently, the perception layer must employ architectures optimized for data efficiency and hardware constraints, such as self-supervised learning methods like the Joint Embedding Predictive Architecture (JEPA) or systems designed for energy efficiency, such as Spiking Neural Networks (SNNs).

The HRM's architecture itself provides structural guidance for sensor integration. It is explicitly modeled on the hierarchical and multi-timescale processing of the human brain, conceptually implementing System 1 (rapid, low-level intuition) and System 2 (slow, abstract planning) thinking.<sup>2</sup> This dual structure must be mirrored by the sensory input pathways, requiring the fusion system to produce both low-latency, real-time feedback for System 1 and high-level, predictive context for System 2.

Furthermore, HRM's efficacy is proven in its ability to achieve exceptional performance on complex inductive tasks (like ARC-AGI) and challenging symbolic tree-search puzzles

(Sudoku-Extreme, Maze-Hard) without resorting to brittle, high-latency Chain-of-Thought (CoT) techniques.<sup>2</sup> This capability validates its capacity for internal, non-verbalized reasoning, making it an ideal recipient for high-density, non-semantic, raw sensory embeddings, as it does not require the sensory output to be pre-converted into human-readable semantic tokens.

---

## II. Architectures for Non-Semantic Multi-Modal Data Fusion

Integrating raw, heterogeneous sensor streams (visual, audio, kinetic) into a coherent, low-latency representation requires a sophisticated approach to fusion. A systematic evaluation of sensor fusion strategies is necessary to select a method that is both highly accurate and computationally aligned with the HRM's efficiency mandate.

### 2.1. A Classification of Sensor Fusion Levels and Challenges

Sensor fusion methodologies are traditionally categorized based on the stage at which data from different modalities is combined.

#### **Fusion Taxonomy and Limitations:**

1. **Feature Level Fusion (Early Fusion):** This involves combining raw data or extracted features (e.g., mean value, standard deviation, spectral energy) from time and frequency domains before feeding them into a classifier.<sup>9</sup> While providing deep integration, this approach is often brittle when dealing with fundamentally disparate data types, such as combining high-dimensional video pixels with low-dimensional IMU vectors.
2. **Decision Level Fusion (Late Fusion):** This method uses modality-specific classifiers to produce individual decisions or probability distributions, which are then merged using techniques like Bagging, Sum, Product, or Maximum rules to form a final classification.<sup>9</sup> While robust to individual sensor failures, Late Fusion sacrifices the opportunity for emergent contextual learning that occurs when data is combined earlier.
3. **Deep Fusion (Integration Layer):** This involves integrating data within the internal layers of a deep neural network, allowing the model to automatically extract relevant features and discover latent relationships. This often yields superior consistency and generalizability compared to systems that rely on manual feature extraction.<sup>9</sup>

The primary difficulty inherent in the user's sensor suite (Camera, Mic, IMU/GPS) is **heterogeneity**. Combining heterogeneous sensors has been shown to yield better average performance in activity recognition compared to fusing homogeneous sensors, as complementary information compensates for the limitations of any single stream.<sup>9</sup> Given the goal of automatic feature extraction and robust context derivation, the architecture must

implement a form of Deep Fusion, but implemented using highly efficient models tailored to handle raw, non-semantic data streams and meet the hardware constraints. The limitations of traditional manual feature extraction—which is labor-intensive and lacks generalizability—further necessitate an approach that performs automatic, consistent, and scalable feature extraction.<sup>9</sup>

## 2.2. Predictive World Modeling: Joint Embedding Predictive Architectures (JEPA)

The Joint Embedding Predictive Architecture (JEPA) presents a compelling solution for the abstract planning requirements of the HRM. JEPA is fundamentally designed as a predictive world model that moves beyond passive observation.<sup>1</sup> In the context of embodied AI, JEPA learns causality and physics by being trained to predict masked features within a data modality or across different modalities. This self-supervised training paradigm achieves high data efficiency and open-vocabulary generalization.<sup>10</sup>

JEPA inherently supports the embodied system requirements by focusing on **Active Egomotion** (where the agent actively configures or moves sensors) and **Active Agency** (where the agent's actions influence the resulting sensory streams).<sup>1</sup> By learning on experienced data—the fundamental difference separating human training from internet-trained AI<sup>1</sup>—JEPA constructs low-dimensional, high-fidelity world state embeddings that capture temporal and spatial causality.

The mechanism of fusion within JEPA-T, for example, utilizes cross-attention after the feature predictor for conditional denoising, combined with objective-level alignment during training.<sup>10</sup> This approach strikes an optimal balance between conditioning strength (allowing modalities to influence each other deeply) and backbone generality (avoiding task-specific brittleness). The resulting architecture serves as the essential predictive grounding layer, generating the required cognitive scaffolding for the HRM's System 2 component. Specifically, JEPA can provide the HRM's High-Level Recurrent Module (HL-RM) with crucial input regarding potential outcomes and long-term consequences, allowing the HL-RM to perform true abstract planning, rather than mere reactive processing.

## 2.3. Energy-Efficient Processing: Spiking Neural Networks (SNNs)

To address the mandate for computational efficiency and real-time response, Spiking Neural Networks (SNNs) are indispensable. Conventional Artificial Neural Networks (ANNs) used for multimodal fusion are often constrained by high computational complexity, memory consumption, and significant energy demands, particularly in resource-constrained or mobile scenarios.<sup>11</sup> SNNs overcome these limitations by operating based on sparse, event-driven spike trains, which are inherently more energy-efficient and offer superior performance in

terms of latency.<sup>11</sup>

SNNs are ideally suited for processing dynamic, event-based data streams generated by the proposed sensor suite. For instance, they naturally integrate with event cameras, which generate data asynchronously only when pixel intensity changes, and dynamic IMU data (accelerometer/gyroscope).<sup>11</sup> An advanced SNN framework for multimodal action recognition utilizes distinct backbones for different modalities—such as an **SNN-based Mamba** for event camera data and a **Spiking Graph Convolutional Network (SGN)** for structured data like skeleton or spatial inputs.<sup>11</sup>

The fusion mechanism within an SNN framework employs a **discretized information bottleneck mechanism**.<sup>11</sup> This method is highly effective for compressing information and balancing the preservation of modality-specific semantics while ensuring efficient information processing. Experiments confirm that this SNN-driven framework delivers superior performance in both recognition accuracy and energy efficiency.<sup>11</sup>

The inherent design of SNNs—optimized for low-latency, event-driven processing—makes them the superior choice for the system's rapid, reactive components. The Low-Level Recurrent Module (LL-RM) of the HRM is dedicated to rapid, detailed computation (System 1).<sup>2</sup> By serving as the front-end for the LL-RM, SNNs ensure that the system can process immediate, non-semantic changes in the environment, such as a sudden acoustic event (microphone spike) or rapid kinematic change (IMU spike), generating features suitable for immediate, low-latency action execution.

Table 1 summarizes the comparative evaluation of these architectures against the constraints of the embodied reasoning system.

Table 1: Comparative Merits of Advanced Sensor Fusion Architectures for Embodied Systems

Architecture	Primary Fusion Level	Key Advantage for Raw Data	Alignment with Embodied Cognition	Energy/Compute Efficiency	HRM Component Target
Traditional Deep Fusion (ANN)	Feature/Deep	High feature complexity extraction	Requires extensive supervised context	Low (High compute/memory) <sup>11</sup>	N/A (Fails Efficiency Test)
Joint Embedding Predictive Architecture (JEPA)	Deep (Self-Supervised)	Excellent for learning predictive world models	Explicitly supports Active Agency/Egomotion <sup>1</sup>	Moderate/High (Data efficient)	High-Level Recurrent Module (HL-RM)
Spiking Neural Networks (SNN)	Early/Feature (Event-Based)	Superior temporal resolution, energy minimization	Biologically plausible, real-time response	High (Low latency, low energy) <sup>11</sup>	Low-Level Recurrent Module (LL-RM)

---

## III. The Hierarchical Reasoning Model (HRM) in Depth

The Hierarchical Reasoning Model (HRM) forms the irreplaceable core of the proposed system. Its design, inspired by cognitive neuroscience, provides the necessary structure to manage planning and execution across multiple timescales while maintaining unparalleled computational efficiency.

### 3.1. Neuroscience Foundations: System 1/System 2 Thinking and Temporal Separation

The architectural foundation of the HRM is rooted in the hierarchical and multi-time scale processing observed in the human brain.<sup>2</sup> This design specifically implements the functional segregation known in cognitive theory as System 1 (intuitive, fast) and System 2 (deliberative, slow) thinking, implemented through two coupled recurrent modules.<sup>6</sup>

The 27M-parameter model achieves significant computational depth—a capability often reserved for models many orders of magnitude larger—by relying on the interdependence of these two modules <sup>2</sup>:

1. **High-Level Recurrent Module (HL-RM):** This module manages slow, abstract planning and long-term goal optimization.<sup>2</sup> Its function is analogous to System 2, focusing on strategic sequence generation and metacognitive oversight.
2. **Low-Level Recurrent Module (LL-RM):** This module handles rapid, detailed computations and immediate action execution.<sup>2</sup> It functions as System 1, responsible for the necessary high-frequency data processing required for interacting with the immediate environment.

The HRM operates through iterative refinement over short "thinking bursts." During each burst, the model produces a work-in-progress prediction and a self-regulatory "halt or continue" score, demonstrating internal processes akin to self-awareness and self-correction.<sup>7</sup>

### 3.2. Performance Analysis: Computational Depth Without CoT

The performance metrics of the HRM underscore its transformative potential. Despite its minimal size and training requirements (only about 1,000 training examples and no pre-training), the HRM surpasses state-of-the-art Chain-of-Thought (CoT) models on inductive benchmarks like the Abstraction and Reasoning Corpus (ARC-AGI), achieving 40.3% accuracy.<sup>6</sup> Crucially, it demonstrates near-perfect accuracy on tasks requiring complex symbolic reasoning, such as Sudoku-Extreme and optimal pathfinding in large mazes, tasks

where CoT models fail completely.<sup>6</sup>

The fundamental benefit here is that the HRM executes sequential reasoning tasks in a single forward pass without the explicit supervision of the intermediate process.<sup>2</sup> CoT techniques, conversely, suffer from brittle task decomposition and high latency.<sup>2</sup> The HRM's capacity for deep, non-verbalized internal reasoning makes it uniquely suitable for integrating non-semantic, raw sensory embeddings, as it eliminates the need to translate the environmental state into an intermediate, human-legible, symbolic format.

### 3.3. Theoretical Extensions: Aligning HRM with Cognitive Architecture

For the HRM to evolve into a sapient agent with self-directed agency, its recurrent modules must be framed within broader cognitive architectures. The HL-RM, dedicated to abstract planning and long-term goal optimization, aligns strongly with the concept of a **Global Workspace Architecture**.<sup>3</sup> The Global Workspace acts as a central integrative bottleneck in consciousness theory, synthesizing competitive broadcasts from various specialized modules into a unified, coherent representation.<sup>12</sup> In the integrated system, the HL-RM would synthesize its internal abstract plan with the dynamically updated representation of the environmental state provided by the sensor fusion layer.

The development of sapience requires the agent to track its cognitive states and adjust its reasoning accordingly, a process known as **computational metacognition**.<sup>3</sup> While the existence of phenomenal consciousness (subjective metacognitive feelings like the "Aha!" moment) is not strictly required for goal discovery and optimization, the functional outcome of these feelings must be approximated.<sup>3</sup> HRM can achieve this through symbolic self-representation, pattern recognition, and predictive modeling.<sup>3</sup>

The LL-RM plays an instrumental role in this architectural context. While defined for rapid action execution, the LL-RM must also function as the essential sensory pre-processor. High-frequency sensor streams cannot be directly fed into the HL-RM without overwhelming the system. The LL-RM, processing rapid, detailed, real-time sensory data, must filter, prioritize, and convert these sensory inputs into meaningful "broadcasts" before transmitting them to the HL-RM's Global Workspace.<sup>12</sup> This establishes the LL-RM as the System 1 filtering mechanism, performing immediate fusion and context aggregation (sensory coherence filtering) to generate a critical, synthesized environmental representation for the System 2 planning module.

---

## IV. Synthesis: Proposed Integration Modalities and Feature Representation

The most critical engineering challenge in integrating the SNN/JEPA perception layer with the HRM core lies in defining the interface: how to map non-semantic features from the raw input

stream into the specialized, dual-timescale recurrent architecture of the HRM.

## 4.1. The Critical Interface: Mapping Non-Semantic Features to Hierarchical Processing

The primary hurdle at this interface is **recoding**.<sup>13</sup> The disparate raw inputs (time-series vectors from IMU, event spikes from camera, spectrograms from audio) must be translated into a common feature vector format that is legible to the HRM's recurrent modules while preserving the integrity of their temporal and spatial relationships. The architecture must ensure the maintained hierarchical progression from sensory data representation to abstract planning representations.<sup>4</sup>

Effective multisensory integration demands that the system overcome computational issues such as statistical inference problems to combine sensory inputs into unitary percepts, ultimately building coherent representations of the environment.<sup>13</sup> For example, the combined spatial context derived from IMU/GPS must be seamlessly fused with the processed visual and auditory data to form a coherent audiovisual spatial representation, similar to how the human brain integrates these modalities.<sup>13</sup>

To meet the architectural needs of the HRM's dual nature, the sensor fusion layer must generate two functionally distinct output streams:

1. **Rapid, High-Detail Stream ( $\mathbf{F}_{\text{rapid}}$ ):** This stream must deliver event-based, low-latency sensory updates, ideally spike frequencies or compressed time-series features. This stream is required by the LL-RM for immediate reaction and fine-tuning. SNNs are the optimal source for this output.<sup>11</sup>
2. **Slow, Abstract Stream ( $\mathbf{E}_{\text{plan}}$ ):** This stream must deliver the predictive context and aggregated world state necessary for abstract planning. This high-level state vector is best provided by JEPA's world model embedding.<sup>1</sup>

## 4.2. Integration Proposal 1: Late Fusion via Abstracted Features (The Low-Risk Baseline)

The simplest integration strategy involves generating a singular, high-level feature vector from the entire sensor front-end and feeding it into the HRM core. In this approach, a single JEPA model, trained on all modalities, generates a compressed feature vector ( $\mathbf{F}_{\text{env}}$ ) representing the current environmental state. This vector is then concatenated with the HRM's previous hidden state ( $\mathbf{h}_{t-1}$ ) and input into both the HL-RM and LL-RM at the beginning of each reasoning burst.

While this approach is advantageous due to its simplicity and preservation of the HRM core architecture integrity, it is highly constrained. It places the entire burden of temporal dependency encoding and dual-timescale feature extraction onto the pre-processing JEPA

layer. This risks the abstraction layer losing the crucial, high-frequency, immediate detail necessary for the LL-RM’s rapid computations.<sup>2</sup> This method is suitable only as a low-risk baseline proof-of-concept.

### 4.3. Integration Proposal 2: Early Fusion within the HRM Low-Level Module (The High-Performance Hybrid)

The optimal, high-performance strategy requires a direct, specialized coupling of the perception architecture to the HRM’s recurrent modules, leveraging the unique temporal and efficiency benefits of SNN and JEPA.

#### Architectural Specialization:

- **JEPA Backbone (HL-RM Input):** JEPA provides the global, predictive context embedding ( $\mathbf{E}_{\text{plan}}$ ). This embedding is derived from self-supervised reconstruction across longer time scales (e.g., temporal changes over several seconds) and informs the HL-RM’s abstract planning module.<sup>1</sup>
- **SNN Backbone (LL-RM Input):** SNNs, utilizing specialized encoders like SNN-Mamba for visual data and SGN for IMU/spatial data <sup>11</sup>, generate low-latency, time-series feature streams ( $\mathbf{F}_{\text{rapid}}$ ). These non-semantic feature streams are input directly into the LL-RM.

**LL-RM as Fusion Processor:** The LL-RM is adapted to function as an Early/Deep Fusion processor for rapid inputs. By having dedicated input gates for the  $\mathbf{F}_{\text{rapid}}$  streams, the LL-RM can perform rapid, detailed computations based on instantaneous environmental changes (the System 1 reflex).<sup>2</sup> This dedicated pathway ensures that immediate, safety-critical responses—such as abrupt evasive action triggered by a collision event detected via the IMU and event camera—can execute rapidly without waiting for the slow, abstract planning cycle of the HL-RM. SNNs are inherently designed to provide the low-latency, event-driven features necessary for this immediate reflexivity.<sup>11</sup> Therefore, feeding SNN outputs directly into the LL-RM maintains the system’s necessary rapid error correction capability, a hallmark of effective embodied agency.

Table 2 formalizes the proposed mapping between the cognitive structures of the HRM and the corresponding sensory inputs.

Table 2: Architectural Mapping: Integrating Sensor Input to HRM Timescales

HRM Component	Processing Timescale	Proposed Sensor Input Source	Required Data Representation	Function in Integrated Model
High-Level Recurrent Module (HL-RM)	Slow, abstract (System 2)	JEPA Predictive World State ( $\mathbf{E}_{\text{plan}}$ ) and Metacognitive Feedback from	Abstract planning tokens, long-term goal states, belief representations <sup>2</sup>	Task decomposition, iterative refinement, causal sequence generation <sup>6</sup>

		LL-RM		
<b>Low-Level Recurrent Module (LL-RM)</b>	Rapid, detailed (System 1)	SNN Rapid Feature Streams ( $\mathbf{F}_{\text{rapid}}$ ); Filtered $\mathbf{E}_{\text{plan}}$	Non-semantic embeddings, Compressed time-series, Event-based spikes <sup>11</sup>	Early fusion of raw features, local environment prediction, immediate action execution, sensory coherence filtering <sup>2</sup>

## V. Architectural Blueprint and Implementation Challenges

The optimal architecture marries the energy efficiency and low latency of SNNs with the predictive power of JEPA, using the HRM’s dual-recurrent structure as the integrative framework.

### 5.1. The Optimal Hybrid Model: SNN/JEPA Front-End to Dual-Recurrent Core

The proposed system is structured as a three-stage processing pipeline:

**1. Modality-Specific Encoding:** Raw sensor streams are processed by specialized encoders tailored to the data type and the required efficiency.

- IMU (Accelerometer/Gyroscope) and GPS: Encoded using a Spiking Graph Network (SGN).<sup>11</sup> SGNs efficiently process the structural and temporal dependencies inherent in spatial and kinematic data.
- Camera (Event-Based or Standard): Processed via an SNN-based Mamba architecture for high-efficiency visual event handling<sup>11</sup> and fed into the JEPA visual encoder.
- Microphone (Audio): Processed using biologically plausible Spiking RNNs optimized for raw spectrogram-level input<sup>4</sup>, converting auditory waveforms into spike trains.

**2. Dual Fusion Pathway:** This stage separates the features into the rapid and abstract streams necessary for the HRM.

- **Rapid Path ( $\mathbf{F}_{\text{rapid}}$ ):** Spiking feature outputs from the SNN encoders (SGN, SNN-Mamba, Spiking RNN) are aggregated, converted (if necessary, though ideally remaining in a spiking or near-spike format), and fed directly into dedicated input gates within the **LL-RM** for System 1 processing.
- **Abstract Path ( $\mathbf{E}_{\text{plan}}$ ):** The feature representations from all modalities inform a global **JEPA World Model**. This model is trained using cross-attention and

self-supervision to predict environmental dynamics and causality.<sup>1</sup> The resulting predictive embedding ( $\mathbf{E}_{\text{plan}}$ ) is fed into the **HL-RM**.

**3. HRM Core Processing:** The HL-RM updates its long-term plan based on the abstracted environmental state ( $\mathbf{E}_{\text{plan}}$ ) and broadcasts necessary plan hypotheses (analogous to the Global Workspace bottleneck<sup>12</sup>). The LL-RM receives these hypotheses, executes the detailed computations based on the rapid sensory updates ( $\mathbf{F}_{\text{rapid}}$ ), refines the prediction, and generates internal self-regulatory feedback (including the "halt/continue" score<sup>7</sup>) back to the HL-RM.

Table 3 provides a detailed implementation map, specifying the required architecture for each sensor to achieve the required dual functionality.

Table 3: Implementation Blueprint: Sensory Modality to Architectural Component Mapping

Sensor Modality	Input Data Type	Primary Encoder Architecture	Primary Feature Stream	Target HRM Component	Rationale for Component Selection
Camera (Event-Based)	Spiking/Raw Pixel	SNN-Mamba <sup>11</sup> / JEPA Encoder	$\mathbf{F}_{\text{rapid}}$ (Visual Events), $\mathbf{E}_{\text{plan}}$ (Visual Context)	LL-RM and HL-RM	SNN-Mamba is optimized for event data efficiency <sup>11</sup> ; JEPA provides the necessary predictive world model context. <sup>1</sup>
IMU (Acc/Gyro) & GPS	Raw Time-Series Vectors	Spiking Graph Net (SGN) <sup>11</sup> / JEPA Encoder	$\mathbf{F}_{\text{rapid}}$ (Kinesthetic/Spatial State)	LL-RM	SGN handles structural/spatial dependencies efficiently <sup>11</sup> ; IMU is critical for immediate reflexes and ego-motion tracking.
Microphone (Audio)	Spectrogram/Waveform	Spiking RNN (Biologically Plausible <sup>4</sup> ) / JEPA Encoder	$\mathbf{F}_{\text{rapid}}$ (Acoustic Events)	LL-RM	RNNs process raw auditory data hierarchically <sup>4</sup> ; feeds into LL-RM for rapid auditory identification.

Fused Output Layer	Abstract/Predictive State	JEPA Fusion Layer (Cross-Attention) <sup>10</sup>	$\mathbf{E}_{plan}$ (World Model)	HL-RM	JEPA is specifically designed to generate unified, predictive embeddings based on experienced data, crucial for System 2 planning. <sup>1</sup>
--------------------	---------------------------	---	-----------------------------------	-------	---

## 5.2. Training Strategies: Leveraging Self-Supervision and Energy Minimization

The training process must be carefully structured to maximize the advantages of both the data-efficient HRM core and the energy-efficient front-end.

**Decoupled Pre-training:** The SNN/JEPA front-end should undergo extensive pre-training using self-supervised methods. This pre-training must leverage large volumes of active environmental interaction data (simulated and real-world egomotion logs) to learn robust non-semantic embeddings that capture causal dependencies.<sup>1</sup> This decoupling allows the HRM core, which requires minimal training data for abstract reasoning (approximately 1,000 samples<sup>8</sup>), to specialize exclusively in high-level problem-solving, dramatically accelerating the path to performance.

**Joint Objective Function:** The final joint fine-tuning process must incorporate two competing optimization objectives:

1. **Reasoning Loss:** Standard loss functions focused on maximizing performance on goal-oriented tasks (e.g., pathfinding, complex problem-solving<sup>8</sup>).
2. **Energy/Latency Loss:** To fully capitalize on the SNN's inherent energy efficiency<sup>11</sup>, energy consumption must be introduced as a penalty term in the loss function. This mechanism encourages the SNN front-end to develop sparser, more efficient spiking activity, ensuring the system remains lightweight and maintainable on consumer-grade hardware.<sup>6</sup>

## 5.3. Enabling Sapient Agency: Designing Feedback Loops

The combination of experience-driven input (Active Agency<sup>1</sup>) and the HL-RM's abstract planning capabilities<sup>2</sup> provides the infrastructure necessary for computational metacognition

and goal optimization—the functional elements of sapience.<sup>3</sup>

A crucial component is the **Causal Feedback Loop**. The action sequence executed by the LL-RM must immediately influence the environment, which is registered by the sensor input streams, thus closing the active agency loop.<sup>1</sup> The discrepancy between the JEPA World Model's prediction ( $\mathbf{E}_{\text{plan}}$ ) and the actual sensory input following an action provides a high-fidelity internal error signal. This prediction error is essential for guiding the HL-RM's abstract planning decisions and continuously refining the internal world model.

## 5.4. Hardware Constraints and Efficiency: Maintaining the Lightweight Advantage

Although the HRM core is remarkably efficient (CPU-deployable<sup>6</sup>), the sensor fusion layer can rapidly become the system's computational bottleneck. The architecture mitigates this by assigning time-critical and detailed sensory processing to SNNs and utilizing JEPA, a model known for its high data efficiency, for abstract planning.<sup>10</sup>

A critical engineering constraint to manage is **Temporal Alignment**. The HRM relies on fixed, iterative "thinking bursts"<sup>7</sup> governed by its recurrent update cycles. In contrast, raw sensor data, particularly from event cameras and IMU, operates asynchronously in continuous time, generating data spikes or updates irregularly.<sup>11</sup> A synchronization and quantization layer is mandatory at the interface between the continuous SNN outputs ( $\mathbf{F}_{\text{rapid}}$ ) and the discrete updates of the HRM. Failure to address this temporal misalignment would introduce noise and destroy the necessary coherence required for accurate multisensory spatial representation.<sup>13</sup>

The final hardware mapping benefits from this duality: SNN components are ideal for deployment on optimized parallel processors or emerging neuromorphic hardware to maximize energy savings, while the JEPA aggregation layer and the HRM core can remain effectively operational on standard high-end consumer CPUs, thus preserving the original lightweight accessibility advantage of the Sapient AI model.<sup>6</sup>

---

## VI. Conclusion and Recommendations

The integration of non-semantic, raw multi-modal sensor fusion with the Hierarchical Reasoning Model constitutes a significant advancement toward creating computationally efficient, embodied general-purpose reasoning systems.

The analysis confirms that standard late-fusion or traditional ANN-based deep fusion methods would fundamentally compromise the efficiency mandate imposed by the 27M-parameter HRM core. The superior architectural blueprint, the High-Performance Hybrid Model, requires a dual-pathway front-end:

1. **Spiking Neural Networks (SNNs):** Serve as the Low-Level Recurrent Module's

dedicated perception engine, providing **low-latency, high-detail feature streams ( $\mathbf{F}_{\text{rapid}}$ ) for immediate System 1 reflexivity** and rapid computation, utilizing energy-efficient SNN-Mamba and SGN backbones.<sup>11</sup>

2. **Joint Embedding Predictive Architecture (JEPA):** Acts as the high-level predictive world model, delivering **abstract, causal state embeddings ( $\mathbf{E}_{\text{plan}}$ ) to inform the High-Level Recurrent Module's System 2 planning.**<sup>1</sup>

This synthesis ensures that the system processes environmental input based on experienced phenomena (Active Agency), rather than passive observation. The LL-RM processes immediate sensor data, filtering it into coherent percepts before broadcasting a unified environmental state to the HL-RM, which manages abstract planning and computational metacognition.<sup>3</sup>

## Recommendations for Future Development:

1. **Prioritize Temporal Synchronization:** Dedicated research must focus on the synchronization layer between the continuous-time SNN output streams and the discrete "thinking bursts" of the HRM, as temporal misalignment represents a critical point of failure for multisensory coherence.<sup>13</sup>
2. **Optimize Joint Loss for Energy:** Implement and empirically evaluate the effect of the energy consumption penalty term in the joint objective function to rigorously quantify and maintain the computational efficiency advantage inherent in the SNN components.<sup>11</sup>
3. **Validate Active Agency Loop:** Focus training validation metrics on the accuracy of the JEPA predictive model against post-action sensory feedback to confirm the system's ability to learn and refine the causal relationships essential for self-directed agency and goal optimization.<sup>1</sup>

## Works cited

1. Deep Dive into Yann LeCun's JEPA - Rohit Bandaru, accessed October 29, 2025, <https://rohitbandaru.github.io/blog/JEPA-Deep-Dive/>
2. Hierarchical Reasoning Model - arXiv, accessed October 29, 2025, <https://arxiv.org/html/2506.21734v1>
3. Recognizing and enhancing sapient agency within AIs: a free will perspective, accessed October 29, 2025, [https://www.researchgate.net/publication/395125021\\_Recognizing\\_and\\_enhancing\\_sapient\\_agency\\_within\\_AIs\\_a\\_free\\_will\\_perspective](https://www.researchgate.net/publication/395125021_Recognizing_and_enhancing_sapient_agency_within_AIs_a_free_will_perspective)
4. Parallel hierarchical encoding of linguistic representations in the human auditory cortex and recurrent automatic speech recognition systems - PMC - NIH, accessed October 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11838305/>
5. Non-Semantic Multimodal Fusion for Predicting Segment Access Frequency in Lecture Archives - ResearchGate, accessed October 29, 2025, [https://www.researchgate.net/publication/394129001\\_Non-Semantic\\_Multimodal](https://www.researchgate.net/publication/394129001_Non-Semantic_Multimodal)

- [\\_Fusion\\_for\\_Predicting\\_Segment\\_Access\\_Frequency\\_in\\_Lecture\\_Archives](#)
6. HRM. The Hierarchical Reasoning Model: What It Is, and More ..., accessed October 29, 2025, <https://shellypalmer.com/2025/09/hrm-the-hierarchical-reasoning-model-what-it-is-and-more-importantly-what-it-isnt/>
  7. The Hidden Drivers of HRM's Performance on ARC-AGI, accessed October 29, 2025, <https://arcprize.org/blog/hrm-analysis>
  8. Hierarchical Reasoning Model - arXiv, accessed October 29, 2025, <https://arxiv.org/html/2506.21734v3>
  9. Multi-Sensor Fusion for Activity Recognition—A Survey - MDPI, accessed October 29, 2025, <https://www.mdpi.com/1424-8220/19/17/3808>
  10. JEPA-T: Joint-Embedding Predictive Architecture with Text Fusion for Image Generation, accessed October 29, 2025, <https://arxiv.org/html/2510.00974v1>
  11. SNN-Driven Multimodal Human Action Recognition via Event Camera and Skeleton Data Fusion Identify applicable funding agency here. If none, delete this. - arXiv, accessed October 29, 2025, <https://arxiv.org/html/2502.13385v1>
  12. (PDF) The Latent Cartesian: Towards a Computational Instantiation of Self-Awareness in Large Language Models (An what then? A multidisciplinary approach) - ResearchGate, accessed October 29, 2025, [https://www.researchgate.net/publication/396161399\\_The\\_Latent\\_Cartesian\\_Towards\\_a\\_Computational\\_Instantiation\\_of\\_Self-Awareness\\_in\\_Large\\_Language\\_Models\\_An\\_what\\_then\\_A\\_multidisciplinary\\_approach](https://www.researchgate.net/publication/396161399_The_Latent_Cartesian_Towards_a_Computational_Instantiation_of_Self-Awareness_in_Large_Language_Models_An_what_then_A_multidisciplinary_approach)
  13. Computational modeling of human multisensory spatial representation by a neural architecture | PLOS One - Research journals, accessed October 29, 2025, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0280987>