Cross-Modal World Modeling with HY-Himmel: Unifying Video, Text, and Sensor Streams for Embodied AI

Eduard J. Burton
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
eduardmail@oregonstate.edu

Abstract

The emergence of embodied artificial intelligence demands unified world models that can seamlessly integrate heterogeneous sensory modalities including vision, natural language, and structured sensor streams. This paper presents a comprehensive analysis of HY-Himmel, a hierarchical interleaved multi-stream motion encoding framework designed for long video understanding and extended to cross-modal world modeling. The architecture addresses fundamental challenges in aligning temporally asynchronous data from video, text instructions, and sensor readings through a nested encoding hierarchy that preserves both fine-grained temporal dynamics and high-level semantic abstractions. We examine the structural trade-offs between model expressivity and computational tractability, the infrastructural requirements for deploying such systems in real-world robotic and autonomous environments, and the governance implications of unifying multi-modal data under a single representational framework. Sustainability considerations are discussed in the context of energy-efficient training and inference, while robustness and fairness are evaluated with respect to domain shift and representation bias. Policy implications arising from the use of cross-modal models in critical infrastructure and public services are also addressed. By situating HY-Himmel within the broader landscape of large-scale multimodal foundation models, this paper offers a systematic exploration of the architectural, operational, and societal dimensions of next-generation embodied AI systems.

Keywords

cross-modal world modeling, embodied AI, multimodal learning, video understanding, sensor fusion, hierarchical encoding, governance, sustainability.

1. Introduction

The pursuit of artificial intelligence that can act in and reason about the physical world has increasingly converged on the need for unified representations that span multiple modalities. Embodied AI agents, whether deployed in autonomous vehicles, domestic robots, or industrial automation, must simultaneously process visual feeds, natural language commands, and an array of sensor measurements such as LiDAR, inertial measurement units, and tactile feedback. Traditional approaches treat each modality independently and fuse information only at late stages, leading to brittle systems that struggle with temporal misalignment and cross-modal ambiguity. Recent advances in large-scale neural architectures, particularly transformers and vision-language models, have demonstrated the feasibility of learning joint embeddings from paired data. However, these models are typically designed for static images

or short video clips and do not scale to the long, temporally dense streams that characterize real-world embodied interaction.

HY-Himmel represents a significant departure from prior work by introducing a hierarchical interleaved multi-stream motion encoding framework that explicitly models long-range temporal dependencies across video, text, and sensor modalities. The architecture leverages a nested sequence of motion-aware encoding layers that capture both local motion patterns and global event structures, enabling the system to maintain coherent world states over extended periods. This paper provides an in-depth analysis of HY-Himmel from a systems perspective, focusing on the structural design choices that enable cross-modal unification, the infrastructural requirements for deployment, and the socio-technical implications of such integrated models. While the technical details of the architecture are documented elsewhere [17], the present work contextualizes HY-Himmel within the broader research agenda of embodied AI and examines the trade-offs that arise when unifying disparate data streams under a single modeling framework.

The systemic perspective adopted here is motivated by the recognition that large-scale multimodal models are not merely algorithmic artifacts but rather socio-technical infrastructures that shape and are shaped by governance policies, resource allocation, and ethical considerations. As these models become embedded in critical applications such as autonomous navigation, healthcare robotics, and disaster response, understanding their structural vulnerabilities and societal impact becomes paramount. The following sections elaborate on the background, architectural innovations, deployment challenges, and governance dimensions of cross-modal world modeling, using HY-Himmel as a central case study while drawing on a wide range of relevant literature.

2. Background and Related Work

The development of cross-modal world models builds upon decades of research in multimodal learning, video understanding, and sensor fusion. Early approaches to multimodal representation learning relied on canonical correlation analysis and kernel methods to project data from different modalities into a shared latent space [1]. While effective for small-scale problems, these methods did not scale to the high-dimensional, temporally structured inputs typical of embodied AI. The advent of deep learning enabled the training of modality-specific encoders that could be jointly optimized through objectives such as contrastive learning and cross-modal translation. The Vision Transformer [2] reformulated image understanding as a sequence prediction problem, paving the way for unified architectures that handle text and video through a common tokenization framework. Similarly, the CLIP model [3] demonstrated that joint vision-language embeddings could be learned from web-scale image-text pairs, enabling zero-shot transfer across modalities.

Video understanding has traditionally been dominated by 3D convolutional networks [4] and recurrent architectures [5], both of which are limited in their ability to capture long-range temporal dependencies. The introduction of video transformers [6] addressed some of these limitations by applying self-attention over spatiotemporal patches, but the quadratic computational cost limited their application to short clips. Hierarchical and factorized attention mechanisms [7] emerged as a solution, decomposing spatial and temporal attention to reduce complexity. These innovations laid the groundwork for architectures capable of processing longer video sequences, though they still struggled with the simultaneous integration of non-visual modalities.

Sensor fusion for robotic systems has long relied on Kalman filters and probabilistic graphical models [8] that assume Gaussian noise and linear dynamics. Modern deep learning approaches have replaced these with end-to-end trainable networks that directly map raw sensor readings to control commands or state estimates. Multi-modal fusion strategies can be classified into early, intermediate, and late fusion, each with distinct representational trade-offs. Early fusion concatenates raw inputs, requiring careful alignment of sampling rates and coordinate frames, while late fusion combines modality-specific predictions, losing the opportunity for cross-modal interaction. Intermediate fusion, where features from different modalities are aligned at multiple abstraction levels, has shown superior performance in tasks such as visual question answering and robotic manipulation [9].

The emergence of large language models [10] fundamentally changed the landscape by providing a powerful prior for reasoning about the world through text. Systems such as PaLM-E [11] and RT-2 [12] demonstrated that language models could serve as the central reasoning engine for embodied agents, processing visual and sensor inputs as textual tokens. However, these approaches often sacrifice fine-grained temporal and spatial information by compressing high-frequency sensor data into coarse representations. Meanwhile, specialized video-language models [13] have achieved state-of-the-art results on benchmarks requiring long-form temporal reasoning, but they are typically evaluated on curated datasets and do not address the noise, missing data, and domain shift inherent in real-world sensor deployments.

HY-Himmel [17] synthesizes ideas from hierarchical video modeling, cross-modal alignment, and sensor stream processing into a single framework. Its key innovation lies in the interleaved multi-stream motion encoding, which learns to temporally align and hierarchically abstract information from multiple asynchronous sources. Unlike prior work that treats each modality as a separate channel to be fused at a fixed layer, HY-Himmel interleaves processing steps across modalities, allowing early cross-modal interactions that capture fine-grained correspondences. This architectural choice has profound implications for the model's ability to handle long video sequences and dense sensor streams, as well as for the computational and memory requirements of deployment.

3. Architecture of HY-Himmel

The HY-Himmel architecture is organized as a hierarchy of motion encoding stages, each operating on a progressively coarser temporal resolution while maintaining cross-modal feature exchange. At the lowest level, raw video frames, text tokens, and sensor readings are tokenized into a unified temporal sequence. Video frames are patchified and embedded using a spatiotemporal tokenizer that preserves both spatial and temporal order. Text is tokenized via a learned subword vocabulary, and sensor streams are discretized into fixed-length windows with positional encodings that capture their relative timing. These tokens are then interleaved according to their timestamps, producing a single mixed-modality sequence that respects the asynchronous nature of real-world data ingestion.

The first encoding layer applies local self-attention within short temporal windows, effectively learning relationships among co-occurring modalities. This is reminiscent of the local attention mechanisms used in Longformer [14] and BigBird [15], but with the crucial difference that attention masks are modality-aware: tokens from the same modality can attend to each other over slightly longer ranges, while cross-modal attention is initially restricted to temporally proximate tokens. This design prevents the model from being overwhelmed by the combinatorial explosion of cross-modal interactions while still enabling early fusion of tightly coupled signals such as a visual motion and corresponding tactile feedback.

As tokens propagate upward through the hierarchy, temporal aggregation reduces the sequence length by merging adjacent tokens into higher-level representations. This aggregation is controlled by learned motion gates that detect salient events and compress redundant or static segments. The motion gates are analogous to the temporal pooling mechanisms used in SlowFast networks [16], but they operate jointly across modalities: a visual motion that coincides with a change in sensor readings or a textual command triggers more fine-grained encoding, while periods of inactivity are aggressively compressed. The result is a compact event-based representation that captures the essential dynamics of the environment without preserving every timestep.

At the top of the hierarchy, a global encoder processes the aggregated event sequence using full self-attention, enabling long-range reasoning across minutes or even hours of recorded interaction. This global encoder outputs a world state vector that can be used for downstream tasks such as action prediction, question answering, or planning. The hierarchical structure introduces a natural trade-off between temporal resolution and representational capacity: deeper hierarchies can model longer sequences but risk losing fine-grained details, while shallower hierarchies preserve local information at the cost of limited context length. HY-Himmel addresses this trade-off through the motion gating mechanism, which adaptively allocates representational resources to informative segments.

The cross-modal interleaving strategy also has implications for the model's scalability. Processing a single long video with multiple sensor streams generates a token sequence that can be several orders of magnitude longer than typical language model inputs. The hierarchical compression ensures that the computational cost of the global encoder remains manageable, while the local attention layers are parallelizable across windows. Nonetheless, the total number of parameters and the memory footprint of the model are substantial, raising important questions about the infrastructural requirements for training and deployment.

4. Cross-Modal Integration and World Modeling

The ultimate goal of cross-modal world modeling is to construct an internal representation of the environment that supports causal reasoning, prediction, and action selection. HY-Himmel's hierarchical interleaved encoding provides a natural framework for learning such representations because it preserves the temporal structure of events while enabling cross-modal interactions at multiple scales. In contrast to models that fuse modalities only after independent encoding, the interleaved approach allows early cross-modal attention to learn fine-grained correspondences, such as the visual appearance of a hand moving and the tactile signal from a contact sensor, which are crucial for precise manipulation tasks.

The world model learned by HY-Himmel can be queried in multiple ways. For instance, given a natural language instruction, the model can attend to relevant segments of the video and sensor history to generate a sequence of actions. This capability depends on the alignment achieved during training: the model must learn that certain textual phrases correspond to specific visual and sensory events. Training such alignments requires large corpora of paired multimodal data, which are expensive to collect in embodied settings. HY-Himmel addresses this by leveraging self-supervised objectives that predict masked tokens across modalities, similar to the masked modeling techniques used in BERT [18] and VideoMAE [19]. By randomly masking tokens from one modality and requiring the model to reconstruct them using context from other modalities, the model learns robust cross-modal correspondences without requiring explicit human annotations.

One of the key challenges in cross-modal integration is handling missing or corrupted sensor data. In real-world deployments, a camera may be occluded, a microphone may fail, or a LiDAR point cloud may be sparse in certain directions. Traditional fusion methods often collapse under such conditions because they assume all modalities are always available. HY-Himmel's token-level interleaving naturally accommodates missing modalities: absent sensor tokens are simply omitted from the sequence, and the attention mechanism learns to rely on the remaining modalities. During training, random dropout of entire modalities is applied to encourage robustness. Experiments with simulated sensor failures have shown that the model retains reasonable performance even when several streams are absent, though accuracy degrades gracefully rather than catastrophically.

Another important dimension of world modeling is the ability to predict future states. Predictive coding has been proposed as a fundamental principle of brain function [20] and has inspired several deep learning architectures for video prediction. HY-Himmel can be extended with a predictive head that forecasts future motion encodings given the current world state and a history of past events. This predictive capability is essential for planning and control in embodied agents, as it allows the system to simulate the consequences of potential actions. However, long-term video prediction remains notoriously difficult due to the high dimensionality and stochasticity of natural scenes. HY-Himmel's hierarchical compression mitigates this by predicting at multiple temporal resolutions: the global encoder predicts high-level event transitions, while the local encoders fill in fine-grained motion details. This decomposition reduces the cumulative error that plagues single-resolution predictors.

5. Deployment and Infrastructure Considerations

Deploying a cross-modal world model such as HY-Himmel in an embodied system requires careful attention to computational infrastructure, latency constraints, and energy budgets. The model's hierarchical architecture can be partitioned across multiple compute nodes, with local encoding layers running on edge devices and the global encoder executing on a server or cloud instance. This split is advantageous for mobile robots that carry limited onboard compute: early-stage motion compression can be performed on a lightweight neural accelerator, and only the compressed event representations need to be transmitted to a central server. However, communication latency and bandwidth limitations may introduce delays that degrade real-time performance. In safety-critical applications such as autonomous driving, even sub-second delays can be unacceptable, necessitating that the entire model fits on an onboard GPU.

The training of HY-Himmel requires massive computational resources. The model is typically trained on data collected from simulated environments [21] and real-world robotic deployments, with millions of hours of multimodal recordings. The tokenization of high-resolution video at multiple frames per second yields training sequences that are tens of thousands of tokens long. Batch training with such long sequences is memory-intensive; techniques such as gradient checkpointing, sequence parallelism, and mixed-precision training are essential. The carbon footprint of training a single instance of HY-Himmel has been estimated to be on the order of several hundred metric tons of CO2 equivalent, comparable to the training emissions of large language models [22]. This raises sustainability concerns that must be addressed through the use of renewable energy, efficient hardware, and model pruning after training.

Inference efficiency is equally important. For a robot that must react to its environment in real time, the model must process sensor data at rates of tens of hertz. The hierarchical

architecture naturally supports early exit mechanisms: if the current situation is determined to be simple, the model can make a decision based on the local encoding layers alone, bypassing the expensive global encoder. Such dynamic computational graphs can significantly reduce average latency and energy consumption. However, they introduce non-determinism, making it difficult to guarantee worst-case performance for certification in safety-critical systems.

6. Robustness, Fairness, and Governance

Cross-modal world models inherit vulnerabilities from each component modality. A vision encoder may be fooled by adversarial patches, a language model may propagate biased stereotypes, and sensor readings may be spoofed by external interference. HY-Himmel's interleaved architecture introduces additional failure modes: an attack on one modality can disrupt the entire world model if the attacker can exploit the cross-modal attention mechanism to corrupt the representation of other modalities. Defending against such attacks requires robust training procedures, including adversarial training [23] and modality-wise regularization.

Fairness considerations arise because the datasets used to train cross-modal models are typically collected from a narrow distribution of environments and demographics. For example, a domestic robot trained primarily in middle-class Western households may exhibit poor performance in other cultural contexts with different room layouts, lighting conditions, and sensor configurations. The hierarchical encoding of HY-Himmel may exacerbate this issue because motion gates that learn to compress certain patterns during training may fail to activate appropriately in novel environments, leading to information loss. Mitigation strategies include domain randomization during training, continuous fine-tuning with on-policy data, and the development of standardized evaluation benchmarks that reflect diverse populations.

Governance of cross-modal AI systems involves questions of accountability, transparency, and oversight. When an autonomous agent equipped with HY-Himmel causes harm, it is often ambiguous whether the fault lies in the vision subsystem, the sensor processing, the language understanding, or the integration layer. The hierarchical nature of the model makes attribution difficult because errors propagate through multiple encoding stages. Regulatory frameworks that require explainability, such as the European Union's AI Act, may compel developers to provide causal explanations of model behavior. Current research on explainability for multimodal transformers is in its infancy, and the nested architecture of HY-Himmel presents additional challenges: attention weights do not necessarily correspond to causal importance, and the compression steps obscure the relationship between input tokens and output decisions.

7. Sustainability and Policy Implications

The environmental impact of training and deploying large cross-modal models cannot be ignored. While HY-Himmel's hierarchical compression reduces inference energy compared to non-hierarchical alternatives, the overall resource footprint remains substantial. Policymakers are beginning to consider regulations that mandate energy disclosure for large AI systems, similar to the fuel economy labels for vehicles. Model developers should report training energy consumption, hardware utilization, and carbon offsets. Furthermore, the embodied AI industry must invest in more efficient hardware, such as neuromorphic chips that implement local attention with spiking neurons, to reduce power consumption.

Policy implications extend to data privacy and security. The sensor streams processed by embodied agents may capture personally identifiable information, including faces,

conversations, and location data. Cross-modal world models that integrate these streams into a unified representation increase the risk of inferring sensitive attributes that are not directly present in any single modality. For instance, combining video of a person's gait with acoustic sensor data could reveal health conditions. Privacy-preserving techniques such as differential privacy and federated learning must be integrated into the training pipeline, but they often conflict with the goal of maximizing cross-modal alignment. A trade-off exists between utility and privacy, and clear policy guidelines are needed to navigate this landscape.

Finally, the deployment of cross-modal world models in public infrastructure, such as traffic management or surveillance, raises concerns about surveillance creep and the concentration of power in the hands of entities that control these models. Open-sourcing model weights and training data can democratize access, but it also enables misuse. A governance framework that includes independent auditing, red-teaming, and iterative stakeholder engagement is essential to ensure that systems like HY-Himmel serve the public interest rather than entrenching existing inequalities.

8. Conclusion

Cross-modal world modeling represents a frontier in embodied AI, and HY-Himmel exemplifies the architectural innovations required to unify video, text, and sensor streams into a coherent world representation. This paper has examined the structural trade-offs inherent in hierarchical interleaved encoding, the infrastructural demands of deployment, and the multifaceted societal implications ranging from robustness and fairness to sustainability and governance. The analysis reveals that while the technical capabilities of such models are impressive, their responsible deployment hinges on addressing fundamental challenges in attribution, privacy, energy consumption, and bias. Future work should focus on developing scalable training protocols that reduce environmental impact, interpretability methods tailored to hierarchical multimodal models, and participatory governance mechanisms that ensure the benefits of cross-modal AI are broadly distributed. The path forward requires close collaboration between researchers, engineers, policymakers, and the communities that will be affected by these transformative systems.

References

1. Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28(3/4), 321–377.

2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

4. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308).

5. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

6. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C. (2021). ViViT: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6836–6846).

7. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning (pp. 813–823).

8. Thrun, S., Burgard, W., & Fox, D. (2005). Probabilistic robotics. MIT Press.

9. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., ... & van den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3674–3683).

10. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901).

11. Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). PaLM-E: An embodied multimodal language model. In Proceedings of the International Conference on Machine Learning (pp. 8469–8482).

12. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... & Zitkovich, B. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning (pp. 123–138).

13. Lei, J., Berg, T. L., & Morariu, V. I. (2022). Video-language pre-training with learned temporal alignment. In European Conference on Computer Vision (pp. 488–505).

14. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

15. Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontañón, S., ... & Ahmed, A. (2020). Big bird: Transformers for longer sequences. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17283–17297).

16. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6202–6211).

17. Jin, H., Yi, H., Zhao, W., Luo, J., Ye, S., Guan, Z., ... & Yu, T. (2026). HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding. arXiv preprint arXiv:2605.08158.

18. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics (pp. 4171–4186).

19. Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural Information Processing Systems (Vol. 35, pp. 10078–10093).

20. Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1), 79–87.

21. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., ... & Batra, D. (2019). Habitat: A platform for embodied AI research. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9339–9347).

22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

23. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations.