# Synthesizing Cross-Modal Decision Policies through Reinforcement Learning Integrating Visual Perception and Large Language Model Tactical Planning

Ethan Ellsworth
Department of Electrical and Computer Engineering, University of New Mexico
e.ellsworth@unm.edu

Scott Norwood
School of Interactive Computing, Georgia Institute of Technology
s.norwood@gatech.edu

Benjamin Pennington
Department of Computer Science, University of Delaware
b.pennington@udel.edu

**Abstract**

The convergence of high-dimensional visual perception and high-level linguistic reasoning represents a frontier in autonomous systems research, particularly concerning the synthesis of robust decision policies. This paper explores the integration of visual sensory inputs with the tactical planning capabilities of Large Language Models (LLMs) within a Reinforcement Learning (RL) framework. While traditional RL excels at low-level motor control and reactive behaviors, it often lacks the semantic depth required for long-horizon strategic navigation in complex, semi-structured environments. Conversely, LLMs provide sophisticated world models and common-sense reasoning but remain fundamentally ungrounded without direct sensory alignment. Our research investigates a hybrid architectural approach where LLMs serve as tactical orchestrators that interpret environmental states conveyed through vision-language encoders, subsequently shaping the reward functions and action spaces for RL agents. We analyze the structural trade-offs inherent in this cross-modal synthesis, focusing on the latency of inference, the stability of the learned policies, and the alignment between symbolic reasoning and physical execution. Beyond the technical mechanics, the study delves into the socio-technical implications of such systems, including their governance, the transparency of cross-modal decision-making, and the long-term sustainability of deploying massive transformer-based models in edge-computing infrastructures. By evaluating these systems through the lens of robustness and fairness, we provide a comprehensive framework for understanding how hybrid cognitive architectures can be scaled responsibly. The findings suggest that while cross-modal integration significantly enhances task generalization, it introduces novel failure modes necessitated by

the stochastic nature of language-based planning, requiring new paradigms for safety-critical deployment.

**Keywords:**
Cross-Modal Synthesis, Reinforcement Learning, Large Language Models, Visual Perception, Tactical Planning, Socio-Technical Systems, Autonomous Policy.

## 1. Introduction

The evolution of autonomous agents has historically been bifurcated into two distinct computational paradigms: the connectionist approach, which emphasizes raw sensory processing and reactive control, and the symbolic approach, which focuses on logic, planning, and high-level reasoning [12]. In recent years, the rapid advancement of deep reinforcement learning has bridged much of this gap, allowing agents to learn complex behaviors directly from high-dimensional inputs [28]. However, a persistent challenge remains in the form of semantic bottlenecking, where agents struggle to generalize beyond their specific training distributions because they lack an abstract understanding of the world. The emergence of Large Language Models (LLMs) offers a potential solution to this limitation by providing a repository of human-acquired knowledge that can be utilized for tactical planning [20]. The integration of these models into decision-making pipelines signifies a shift toward cross-modal intelligence, where visual perception is not merely a trigger for a learned reflex but a prompt for a sophisticated reasoning engine [5].

This paper addresses the synthesis of decision policies that effectively merge the granular, high-frequency requirements of visual-motor control with the abstract, low-frequency requirements of strategic planning. By utilizing reinforcement learning as the connective tissue, we can create agents that are both physically grounded and intellectually capable [19]. This synthesis is not merely a matter of stacking neural layers; it involves a fundamental reimagining of the agentic architecture. We must consider how the latent representations of a visual scene can be translated into a linguistic context that an LLM can process, and conversely, how the high-level directives of an LLM can be distilled into the numerical reward signals that drive RL optimization [31]. The complexity of this interaction necessitates a deep dive into the infrastructure that supports these models, particularly concerning the computational costs and the governance frameworks required to ensure that such agents act in alignment with human values [8].

The motivation for this research stems from the increasing demand for autonomous systems in environments that are both physically demanding and socially complex, such as urban search and rescue, healthcare assistance, and collaborative manufacturing [3]. In these domains, an agent cannot rely solely on a fixed set of visual cues. It must understand the underlying rationale behind its actions and be able to adapt its tactics based on verbal instructions or changing environmental norms [22]. As we move toward more pervasive artificial intelligence, the robustness and fairness of these decision policies become paramount. If a vision-language model misinterprets a cultural nuance or a spatial relationship, the resulting action in the physical world could be catastrophic [14]. Therefore, this study

emphasizes the socio-technical infrastructures that surround the deployment of cross-modal reinforcement learning, arguing that the engineering of these systems must be inextricably linked to their ethical and policy-oriented oversight [27].

## 2. Architectural Frameworks for Cross-Modal Integration

The structural design of a system that integrates visual perception, language-based reasoning, and reinforcement learning must account for disparate temporal and spatial scales. Visual perception typically operates on a frame-by-frame basis, requiring high throughput to capture the dynamics of a physical environment [1]. In contrast, tactical planning via an LLM is a more deliberate process, often involving multi-step reasoning chains that operate at a much slower cadence. The primary architectural challenge lies in creating a synchronization mechanism that allows the high-level plan to inform the low-level policy without introducing debilitating latency [24]. One prevalent approach is the hierarchical reinforcement learning framework, where the LLM functions as a high-level manager that sets sub-goals for a low-level worker agent [11]. The worker agent, trained through standard reinforcement learning techniques, focuses on achieving these sub-goals using visual feedback, while the LLM monitors the broader mission objectives [29].

Within this hierarchical structure, the communication protocol between the visual encoder and the LLM is critical. Simple image captioning is often insufficient for complex tactical planning because it loses the spatial granularity required for precise movement [4]. Instead, modern architectures utilize vision-language models that generate dense embeddings or visual tokens which can be interleaved with text tokens [17]. This allows the LLM to perceive the environment in a way that is natively compatible with its transformer-based architecture [18]. However, the trade-off for this integration is a massive increase in the state-space complexity. Agents must learn to filter out irrelevant visual noise while attending to the specific linguistic constraints provided by the tactical planner [9]. The governance of this information flow is a key area of research, as it determines how much autonomy the low-level agent has to override a high-level plan if it detects an immediate physical threat that the planner has not yet processed [34].

Furthermore, the deployment of these architectures requires a robust computational infrastructure. The sheer size of large language models makes them difficult to run on edge devices, leading to a heavy dependency on cloud-based inference [13]. This introduces a vulnerability in the form of network reliability. If an autonomous vehicle or a robotic assistant loses its connection to the tactical planner, its decision policy must be resilient enough to revert to a safe, purely visual-reactive mode [26]. We argue that a truly robust cross-modal system should feature graceful degradation, where the agent's capabilities scale down according to the availability of its various cognitive modules [16]. This systemic resilience is essential for the long-term sustainability of AI-driven infrastructures, especially as we begin to rely on them for critical societal functions [25].

## 3. Reinforcement Learning and the Grounding Problem

One of the most significant hurdles in artificial intelligence is the grounding problem—the

difficulty of associating abstract symbols with physical realities. In a cross-modal decision policy, this manifests as a gap between the linguistic tactical suggestions and the physical constraints of the agent [7]. For instance, a model might suggest moving an object to a specific location without inherently understanding the torque required for the maneuver or the friction of the surface [33]. Reinforcement learning serves as the primary mechanism for bridging this gap by providing a feedback loop where the agent's actions result in physical consequences [2]. By rewarding the agent not just for completing the task, but for following the tactical nuances of the linguistic planner, we can teach the agent to ground linguistic concepts in sensory experience [21].

The synthesis of these policies often involves a dual-reward structure. The primary reward is task-oriented, while the secondary reward is alignment-oriented, measuring how well the agent's trajectory matches the tactical plan [10]. This approach, however, introduces the risk of reward hacking, where the agent finds a way to satisfy the numerical reward without actually performing the desired behavior in a meaningful way [15]. To mitigate this, researchers are exploring the use of thought-based decision making, where the agent must justify its actions through a latent reasoning process before execution [30]. This relates to the concept of thinking fast and slow in decision models, where the system balances immediate visual response with deliberate linguistic planning [6].

Infrastructure for training such grounded models is also a major concern. The data requirements for reinforcement learning are notoriously high, and when coupled with the need for diverse vision-language interactions, the scale of simulation required becomes immense [32]. We are seeing a move toward high-fidelity digital twins—virtual environments that mirror physical spaces—to provide the necessary training grounds [23]. Yet, even the best simulations suffer from the sim-to-real gap. A policy that learns to navigate a virtual space might fail in a real environment due to lighting changes or unexpected obstacles [15]. The robustness of a cross-modal policy is therefore measured by its ability to maintain tactical coherence even when its visual perception is degraded or noisy. This requires a level of meta-learning, where the agent learns how to learn from its mistakes across both modalities [12].

**4. Tactical Planning and Long-Horizon Governance**

Tactical planning in cross-modal systems is not just about reaching a goal; it is about navigating the complex web of constraints that define human environments. Large language models are uniquely suited for this because they can incorporate social norms, safety protocols, and ethical guidelines into their plans through sophisticated prompting [14]. However, the governance of these plans is a non-trivial task [27]. Unlike traditional deterministic code, the output of a language model is probabilistic and can be unpredictable. When this output is used to drive a reinforcement learning policy, the risk of plan-drift—where the agent gradually moves toward an unsafe state due to a sequence of plausible but slightly flawed tactical decisions—must be addressed [35].

Effective governance requires a multi-layered approach to policy synthesis. At the highest

level, there must be a set of immutable safety constraints that no tactical plan can violate [13]. These are often implemented as shielding mechanisms in the reinforcement learning environment. Below this, the generated plan must be subjected to a rigorous verification process. This might involve using an independent model to critique the proposed plan for potential biases or safety risks [8]. This internal check-and-balance system is crucial for ensuring fairness, particularly when the agent is interacting with diverse human populations [26]. For example, a delivery robot's tactical planner must be governed to ensure it does not prioritize certain regions over others based on biased training data [23].

The policy implications of these systems extend to the regulatory sphere. As cross-modal agents become more integrated into infrastructure, the question of responsibility for failures becomes more complex [5]. Is it the developer of the visual encoder, the provider of the tactical model, or the engineer who designed the reward function? The interdisciplinary nature of these systems demands a new legal framework that accounts for distributed causality in AI decision-making [16]. We propose that the tactical planning module should maintain an audit trail of its reasoning process, allowing human overseers to understand the rationale behind specific actions [31]. This transparency is not just an engineering requirement; it is a fundamental component of social trust in autonomous systems [25].

**5. Sustainability and Deployment Infrastructure**

The deployment of large-scale cross-modal models raises significant questions about environmental and economic sustainability. The energy consumption required to train and run these models is substantial, and as they become more ubiquitous, the aggregate carbon footprint could become a major concern [33]. Engineering more efficient architectures is therefore a priority. This includes the development of specialized hardware, such as neuromorphic chips or energy-efficient processing units, that can handle the parallel processing demands of vision and language at the edge [32]. Furthermore, the use of model distillation—where a smaller student model learns to replicate the behavior of a massive teacher model—is a promising avenue for reducing the computational burden [18].

From an infrastructure perspective, the shift toward decentralized or federated learning could enhance sustainability [6]. Instead of sending all sensory data to a central cloud server, agents could perform local policy updates and share only the learned weights. This would not only reduce bandwidth usage but also enhance privacy, as raw visual data would never leave the local device [21]. However, federated reinforcement learning is notoriously difficult to stabilize, especially when dealing with the non-stationary distributions typical of cross-modal environments [1]. The trade-off between the efficiency of centralized training and the privacy or sustainability of decentralized deployment remains a central tension in the field [10].

Deployment also involves the management of technological debt. As autonomous systems are updated with newer tactical models or better visual encoders, legacy policies might become obsolete or incompatible [22]. A sustainable socio-technical infrastructure must account for the entire lifecycle of the autonomous agent, from initial training to decommissioning [27]. This includes creating modular architectures where individual components can be swapped

out without requiring a total retraining of the entire system [11]. Such modularity promotes robustness by allowing for targeted fixes to specific failure modes, whether they occur in the perception, planning, or execution phase of the decision policy [4].

## 6. Robustness, Fairness, and Public Policy

The ultimate measure of a cross-modal decision policy is its performance in the real world, particularly its resilience in the face of adversarial attacks or unforeseen environmental shifts. Robustness in this context refers to the agent's ability to maintain its tactical integrity when its inputs are compromised [30]. For instance, adversarial patches in the visual field can cause a model to misidentify objects, potentially leading the tactical planner to make dangerous decisions [15]. Ensuring robustness requires a combination of adversarial training, where the agent is exposed to these attacks during the reinforcement learning phase, and formal verification of the reasoning logic [9].

Fairness is equally critical. Cross-modal systems often inherit the biases present in their massive training datasets [20]. A visual perception system might be less accurate at detecting certain demographic groups, or a language model might harbor assumptions about professional roles [35]. When these biases are baked into a decision policy, they can lead to discriminatory outcomes in areas like autonomous policing or healthcare [14]. Addressing this requires a proactive approach to de-biasing at every level of the stack—from the datasets used for visual pre-training to the reward functions that shape final behavior [26]. Public policy must play a role here, setting standards for transparency and accountability that force developers to prioritize fairness over raw performance [8].

The socio-technical implications of these technologies also involve their impact on the labor market and human agency [3]. As agents become more capable of complex tactical planning, they may begin to replace humans in roles that require high-level coordination [25]. This necessitates a broader discussion about the governance of AI in the workplace and the potential for human-AI collaboration [16]. Rather than designing agents that act purely autonomously, we should focus on human-in-the-loop systems where the cross-modal policy acts as an assistant, enhancing human decision-making rather than replacing it [31]. This collaborative approach can mitigate some of the risks associated with fully autonomous policies while leveraging the unique strengths of both human intuition and machine scale [19].

## 7. Case Illustrations and Cross-Domain Comparisons

To illustrate the practical application of these integrated policies, we can look at the domain of urban search and rescue. In this scenario, an autonomous drone must navigate a collapsed building to find survivors [1]. The visual perception module must identify obstacles, structural weaknesses, and human signs of life in low-light, dusty conditions [2]. Simultaneously, the tactical planning module must coordinate with other drones, prioritize areas based on heat maps, and follow safety protocols regarding structural stability [29]. A reinforcement learning agent trained without a tactical planner might find the quickest path but could ignore a high-level command to avoid an unstable zone [12]. The synthesis of these modalities allows

for a policy that is both tactically sound and physically capable [28].

Comparing this to the domain of autonomous driving, we see different structural trade-offs. In driving, the latency requirements are much stricter [24]. A delay of half a second in a tactical decision can be fatal [34]. This necessitates a more tightly coupled architecture where the tactical model's role might be limited to high-level route planning and social interaction, while a more traditional, high-speed reinforcement learning policy handles the immediate maneuvering [13]. These cross-domain comparisons highlight that there is no one-size-fits-all solution for cross-modal synthesis [32]. The optimal architecture depends heavily on the specific temporal and safety constraints of the application domain [17].

In collaborative manufacturing, the focus shifts toward common-sense reasoning [3]. A robot working alongside a human must understand not just the physical task, but the human's intentions [7]. If a human reaches for a tool, the robot's tactical planner should anticipate this and move out of the way or hand the tool to the person [31]. This requires the model to process visual cues as social signals [5]. The decision policy here is not just about efficiency; it is about social robustness—the ability to maintain a productive and safe interaction even when human behavior is unpredictable [22]. This reinforces the idea that cross-modal policies are fundamentally socio-technical artifacts that must be designed with a deep understanding of human psychology and social dynamics [27].

**8. Forward-Looking Perspectives and Research Frontiers**

The future of cross-modal decision policies lies in the development of more world-aware models [4]. Current language models are trained primarily on text, but future iterations will likely be natively multimodal from the start, trained on massive datasets of video, audio, and sensorimotor data [18]. This would eliminate many of the grounding problems we see today, as the model would have an inherent understanding of physical causality [21]. We also anticipate a move toward more interpretable reinforcement learning, where the agent's learned policy can be deconstructed into a series of human-understandable rules or logic gates [11]. This would significantly improve the governance and auditability of these systems [10].

Another frontier is the integration of fast and slow thinking mechanisms more deeply into the neural architecture [6]. Rather than having a separate planner and actor, we might see a single unified model that can dynamically allocate computational resources between quick reactive processing and deep deliberate reasoning depending on the complexity of the situation [30]. This cognitive fluidity would mirror the human brain's ability to switch between autopilot and focused attention [19]. Such an architecture would be inherently more robust and efficient, as it would not waste energy on complex reasoning for simple tasks [20].

Finally, the ethical and legal frameworks governing these systems must evolve in tandem with the technology [3]. We need international standards for the deployment of cross-modal agents in public spaces, ensuring that they are used in ways that are transparent, fair, and sustainable [25]. The goal of our research is to provide the foundational technical and systemic understanding necessary to build these future infrastructures [16]. By synthesizing visual

perception and linguistic tactical planning through reinforcement learning, we are not just building smarter machines; we are creating the building blocks for a new era of intelligent, grounded, and socially responsible autonomous systems [33].

## 9. Conclusion

The synthesis of cross-modal decision policies represents a pivotal moment in the advancement of autonomous systems. By integrating the high-level reasoning of Large Language Models with the grounded, sensory-driven learning of reinforcement learning and visual perception, we can create agents that bridge the gap between abstract thought and physical action. This paper has explored the complex architectural, infrastructural, and socio-technical challenges inherent in this integration. We have argued that while the technical hurdles are significant—ranging from latency and state-space complexity to the sim-to-real gap—the more profound challenges lie in the governance, fairness, and sustainability of these systems.

A robust decision policy is one that not only achieves its physical goals but does so within a framework of safety and ethical alignment. This requires a multi-layered approach to design, incorporating hierarchical structures, internal verification modules, and transparent audit trails. Furthermore, the deployment of these systems must be supported by an infrastructure that prioritizes environmental sustainability and data privacy. As we look toward a future where autonomous agents are increasingly integrated into our daily lives, the interdisciplinary insights provided in this study offer a roadmap for developing AI that is both highly capable and deeply aligned with human values. The journey toward truly cross-modal intelligence is as much an exercise in social engineering as it is in computer science, requiring a commitment to building systems that are as trustworthy as they are intelligent.

## References


1. Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine, 34(6), 26-38.

2. Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. International Conference on Machine Learning, 449-458.

3. Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. Communications of the ACM, 64(7), 58-65.

4. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

5. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

6. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf:

Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

7. Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346.

8. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

9. Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

10. Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 32(11), 1238-1274.

11. Kulkarni, T. D., Narasimhan, K., Saeedi, A., & Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. Advances in Neural Information Processing Systems, 29.

12. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

13. Leslie, D. (2019). Understanding artificial intelligence ethics and safety. The Alan Turing Institute.

14. Liao, Q. V., & Kushlev, K. (2021). Human-centered AI. ACM Interactions, 28(4), 30-35.

15. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

16. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

17. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

18. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32.

20. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.

21. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

22. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477-486.

23. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Jost, J., & Barnes, D. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 145-151.

24. Riedmiller, M., Hafner, R., Lampe, T., Neunert, M., Degrave, J., Wieleba, T., ... & Springenberg, J. T. (2018). Learning by playing-solving sparse reward tasks from scratch. International Conference on Machine Learning, 4344-4353.

25. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

26. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495-504.

27. Simon, H. A. (1996). The Sciences of the Artificial. MIT Press.

28. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140-1144.

29. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.

30. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. IEEE/RSJ International Conference on Intelligent Robots and Systems, 23-30.

31. Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press.

32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

33. Verbeek, P. P. (2011). Moralizing Technology: Understanding and Designing the Morality of Things. University of Chicago Press.

34. Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3-4), 279-292.

35. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.