ReflectiChain: Epistemic Grounding in LLM-Driven World Models for Supply Chain Resilience

Anonymous Authors[1]

Abstract

Global supply chains under geopolitical volatility expose the semantic deficits of traditional OR/RL and the physical ungroundedness of pure LLMs. We introduce REFLECTICHAIN, an agentic framework utilizing a Generative Supply Chain World Model (SC-WM) for resilience planning, formulated as a Constrained POMDP. Using a topology-aware latent dynamics model, the agent performs multi-step "mental rehearsals" to evaluate strategies. Crucially, a Double-Loop Learning mechanism couples test-time reflection-in-action with retrospective reflection-on-action via LoRA updates, enabling autonomous self-evolution. Validated on the multi-agent Semi-Sim benchmark under adversarial policy shocks, REFLECTICHAIN achieves Pareto-optimal compliance, resilience, and profitability. It maintains an 82.3% operability ratio under extreme pressure, proving latent rehearsal is fundamental for approximating Nash Equilibria in complex networks.

1. Introduction

Transnational supply chains enable large-scale forecasting (1; 2), yet quantifying semiconductor resilience remains challenging (3; 4). fragmented "designed in usa, manufactured in east asia" architectures (5) render these systems hypersensitive to geopolitical volatility (6). traditional models fail during "policy black swan" events (7) by neglecting non-linearities driven by epistemic uncertainty and hard policy constraints rather than mere stochastic drift. this necessitates a shift from aleatory inference toward epistemic intelligence (ei) to navigate the "unknown unknowns" of global infrastructure through strategic environmental understanding.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Under review by the 2nd Workshop on Epistemic Intelligence in Machine Learning (EIML@ICML 2026). Do not distribute.

World models (wm), particularly predictive architectures like jepa (8), internalize complex environmental laws, showing promise in embodied ai (9) and state synthesis (10). although modern representation learning captures structured temporal dynamics across heterogeneous signals (11; 12; 13; 14; 15; 16), applying wm to global supply chains faces epistemic bottlenecks. first, supply chains' high-dimensional topology—distinct from 2d pixel grids—causes representation distortion and the curse of dimensionality (17) in vision-centric models. second, pixel-level inference is computationally prohibitive, failing to meet the epistemic efficiency and real-time accountability required for industrial decision-making under uncertainty.

To address these epistemic bottlenecks, we propose an agentic world model framework and introduce REFLECTICHAIN: a unified architecture for resilient and policy-aware supply chain planning. inspired by embodied ai (18), we reformulate supply chain management as a "trajectory planning problem under epistemic constraints." This framework leverages large language models (llms) as semantic interpreters to translate unstructured policy texts into precise mathematical constraints while identifying "unknown unknowns" within policy dynamics. our design integrates advances in multi-agent simulation, semantic reasoning, and logical robustness to construct a self-evolving epistemic system (19; 20; 21; 22; 23; 24). Unlike conventional pixel-based generation, our world model simulates supply chain state trajectories within a latent space (25), maintaining high fidelity with minimal computational overhead. furthermore, we introduce a "rehearse-reflect-correct" mechanism leveraging scaling test-time compute (26). this mechanism endows the system with epistemic introspection, enabling human-like "dream-state" simulations to actively identify knowledge boundaries and iteratively refine decision quality during inference.

We propose REFLECTICHAIN, an agentic world model framework that reformulates supply chain management as trajectory planning under epistemic constraints (18). our architecture leverages llms to translate unstructured policies into mathematical constraints while uncovering "unknown unknowns" within global dynamics (19; 20; 21; 22; 23; 24). by simulating state trajectories in a latent space (25) rather than pixels, REFLECTICHAIN maintains

high fidelity with minimal overhead. crucial to this is a "rehearse-reflect-correct" mechanism driven by scaling test-time compute (26), which endows the system with epistemic introspection to iteratively refine decision quality through "dream-state" simulations during inference.

The main contributions of this paper are summarized as follows:

- **agentic world model for policy-sensitive supply chains:** we introduce the first framework integrating llm-based reasoning with latent world models to encode geopolitical policies as epistemic constraints, enabling autonomous exploration of risk propagation paths.
- **latent test-time reflection algorithm:** we develop efficient latent operators that decouple simulation from intensive video hardware, enabling long-horizon risk rehearsal and non-linear gains in decision quality via scaling inference-time search compute.
- **semi-sim benchmark and counterfactual validation:** we present a high-fidelity semiconductor simulation environment for counterfactual analysis of "black swan" events, demonstrating pareto-optimal profit, resilience, and compliance under unpredictable policy-driven shocks.

2. Related Work

LLMs for Operations and Supply Chain Traditional Operations Research (OR) and Reinforcement Learning (RL) frequently exhibit fragility when confronted with the high-dimensional uncertainty of modern supply chains (27). Large Language Models (LLMs) offer a robust alternative, evolving from zero-shot time-series and multivariate forecasting (28; 29; 30; 31; 32) into autonomous solvers for complex routing and logistics (33; 34; 35; 36; 37). To navigate macro-environmental volatility, recent works integrate LLMs with Knowledge Graphs (KGs) (38) and macroeconomic indices (39; 40; 41) to parse unstructured geopolitical risks. However, these static interpreters lack physical grounding, often suffering from "semantic blindness" during non-stationary "black swan" events. In contrast, our work introduces an adaptive test-time framework seamlessly unifying semantic policy comprehension with grounded physical representations.

Generative World Models and Planning World models enable latent rehearsal and long-horizon planning by approximating environmental dynamics. While mainstream research focuses on computationally heavy pixel-level synthesis for embodied AI and video generation (42; 9; 43), efficient alternatives employ discrete or graph-structured latent spaces (44; 45). Recently, LLM-driven frameworks have coupled LLM planners with world model simulators via tree search methods (46; 47; 48). Yet, visual models remain prohibitive for macro-supply chain graphs, and text-only simulators hallucinate complex cascade dynamics like the bullwhip effect. Our proposed SC-WM resolves this by generating risk trajectories within a compressed graph-embedded latent space, facilitating computationally efficient, high-fidelity policy evaluation.

Reflection and Agentic Self-Evolution Reflection mechanisms facilitate agentic self-evolution through iterative self-criticism. Early single-loop verbal reflection paradigms (49; 50; 51) were subsequently enhanced by multi-agent debate and tool-use (52; 53; 54). The field has since advanced from prompt-level refinement to test-time training (TTT), dynamically updating model weights during deployment (55; 56**?**). Despite this, conventional verbal reflection struggles with the severe delayed credit assignment inherent in macro-supply chains and fails to adapt to non-stationary distribution shifts. ReflectiChain overcomes this through true Double-Loop Learning: operationalizing retrospective hindsight into explicit test-time policy gradients, thereby fundamentally evolving the agent's generative distribution rather than merely appending context to its prompt.

3. Reflective Test-Time Planning via Generative World Models

Figure 1 presents the overall architecture of REFLECTICHAIN. We formalize macro-supply chain management under policy volatility as a Constrained Partially Observable Markov Decision Process (C-POMDP), denoted by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O}, \mathcal{C}, \gamma)$. At each strategic timestep t, the agent receives a multimodal observation $o_t \in \Omega$ (comprising structured inventory states and unstructured geopolitical policy streams). The agent's objective is to find a policy π that maximizes the expected discounted return $\mathbb{E}\left[\sum \gamma^t r_t\right]$ while strictly adhering to non-stationary epistemic constraints $\mathcal{C}_{policy}$ derived from complex international trade regulations.

3.1. The Generative Supply Chain World Model (SC-WM)

To overcome the limitations of black-box models and the high-dimensional topology of supply networks, SC-WM eschews raw pixel or text generation. Inspired by the Joint-Embedding Predictive Architecture (JEPA), we design a topology-aware, compressed latent dynamics model. The environment state is modeled as a dynamic heterogeneous graph $G_t = (\mathcal{V}, \mathcal{E}, X_t)$, where $\mathcal{V}$ represents supply nodes, $\mathcal{E}$ denotes trade dependencies (incorporating weights and flows), and X_t encapsulates node-level operational features

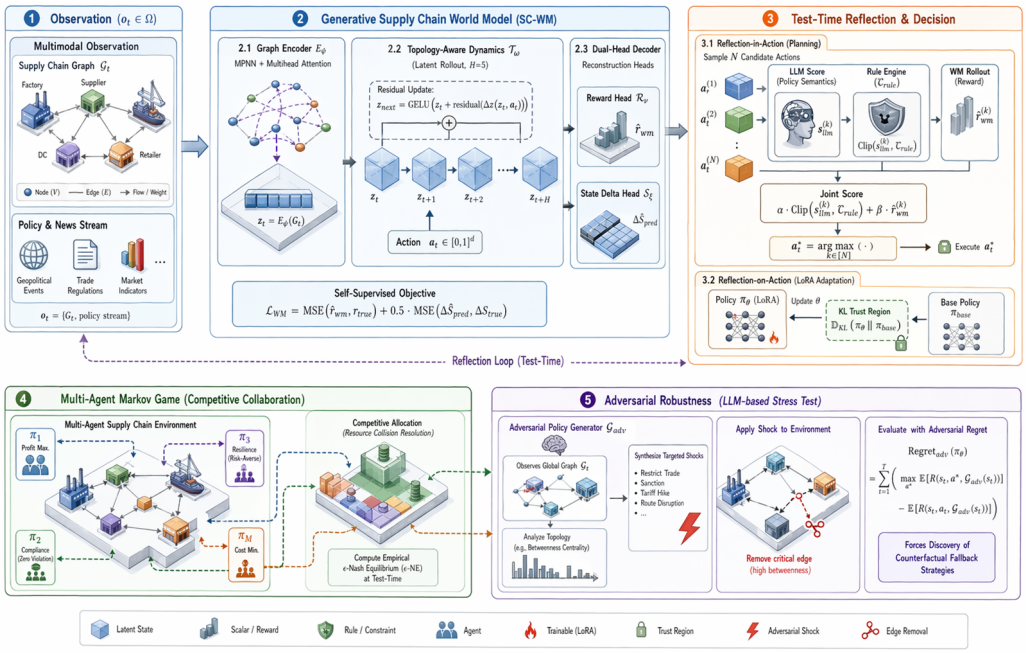

Figure 1: Overall architecture of REFLECTICHAIN. The framework integrates semantic policy understanding, latent world-model-based trajectory rehearsal, and retrospective test-time adaptation for resilient supply chain decision-making.

(e.g., inventory, cash flow).

Architecture & Latent Rollout: SC-WM follows a standard Encoder-Dynamics-Decoder paradigm:

1. **Graph Encoder (E_ψ):** A Message-Passing Neural Network (MPNN) encodes edge features and concatenates them with node features. Utilizing Multihead Attention to dynamically capture critical risk propagation paths, it maps the complex observation graph to a low-dimensional latent state $z_t = E_\psi(G_t)$.

2. **Topology-Aware Dynamics ($\mathcal{T}_\omega$):** Acting as the core simulation engine, it employs residual connections to predict state evolution: $z_{next} = \text{GELU}(z_t + \text{residual}(\Delta z(z_t, a_t)))$. This module supports multi-step latent rollouts (e.g., horizon $H = 5$), enabling the agent to conduct "mental rehearsals" to evaluate long-term action impacts prior to physical execution.

3. **Dual-Head Decoder ($\mathcal{R}_\nu, \mathcal{S}_\xi$):** Reconstructs latent predictions into physical metrics, comprising a Reward Head ($\hat{r}_{wm}$) for comprehensive action returns, and a State Delta Head ($\Delta\hat{S}_{pred}$) for physical state variations.

Self-Supervised Objective: The model is fine-tuned on historical interaction trajectories using a joint loss function that penalizes both reward estimation and physical conservation violations:

$$\mathcal{L}_{WM}(\psi, \omega, \nu, \xi) = \text{MSE}(\hat{r}_{wm}, r_{true}) + 0.5 \cdot \text{MSE}(\Delta\hat{S}_{pred}, \Delta S_{true}) \tag{1}$$

This physical grounding eradicates the circular evaluation bias inherent in purely LLM-based scoring. Furthermore, strictly normalizing the action space (e.g., mapping procurement volumes to $a_t \in [0, 1]$) prevents out-of-distribution (OOD) collapse under extreme scale fluctuations.

3.2. Test-Time Reflection & LoRA Adaptation

During inference, the framework executes **Reflection-in-Action** by scaling test-time compute. For N sampled candidate actions, the optimal action maximizes a joint score:

$$a_t^* = \arg\max_{k \in [N]} \Big(\alpha \cdot \text{Clip}(s_{llm}^{(k)}, \mathcal{C}_{rule}) + \beta \cdot \hat{r}_{wm}^{(k)} \Big) \tag{2}$$

Crucially, to prevent LLM hallucinations and address the "circular evaluation" critique, s_{llm} is strictly bounded by a rule-based physics engine $\mathcal{C}_{rule}$, which enforces deterministic physical constraints (e.g., mass conservation).

For **Reflection-on-Action**, we perform policy gradient updates on the LoRA weights θ. To mitigate catastrophic pol-

icy drift during test-time adaptation—a known instability in continual learning—we introduce a KL-divergence trust-region penalty relative to the base policy π_{base}:

$$\nabla_\theta \mathcal{J}(\theta) \approx \sum_{j \in \mathcal{B}} r^{(j)} \nabla_\theta \log \pi_\theta(a_j \mid o_j) - \eta_{KL} \nabla_\theta \mathbb{D}_{KL}(\pi_\theta \| \pi_{base}) \tag{3}$$

3.3. Multi-Agent Gaming and Adversarial Robustness

Recognizing the decentralized nature of modern supply chains, we extend ReflectiChain into a **Multi-Agent Markov Game**. Let $\Pi = \{\pi_1, \ldots, \pi_M\}$ denote heterogeneous agents with conflicting objectives (e.g., π_1 pursues aggressive profit maximization, while π_2 demands strict zero-compliance violation). At timestep t, SC-WM resolves resource collisions via a competitive allocation mechanism, enabling the computation of an empirical ϵ-Nash Equilibrium (ϵ-NE) during test-time evolution.

To rigorously evaluate system resilience against non-stationary policy shocks ("Black Swan" events), we replace random noise injection with an LLM-based **Adversarial Policy Generator** ($\mathcal{G}_{adv}$). During evaluation, $\mathcal{G}_{adv}$ observes the global graph G_t and synthesizes targeted geopolitical shocks (e.g., severing the edge with the highest betweenness centrality). The evaluation objective shifts from simple reward maximization to minimizing **Adversarial Regret**:

$$\text{Regret}_{adv}(\pi_\theta) = \sum_{t=1}^{T} \Big(\max_{a^*} \mathbb{E}[R(s_t, a^*, \mathcal{G}_{adv}(s_t))] - \mathbb{E}[R(s_t, a_t, \mathcal{G}_{adv}(s_t))] \Big) \tag{4}$$

This adversarial pressure test forces ReflectiChain to discover counterfactual fallback strategies rather than overfitting to historical shock distributions.

4. Experiments

To rigorously assess the effectiveness and robustness of REFLECTICHAIN, we instantiate our framework in SEMI-SIM, a high-fidelity simulation environment described in Appendix A (Evaluation Metrics and Benchmark Construction). Unlike static benchmarks, SEMI-SIM models a 6-node, 7-edge semiconductor supply network governed by epidemic (SIR) risk propagation dynamics and C-POMDP constraints. To expose the fundamental vulnerabilities of baseline methods, the environment is initialized under extreme pressure: strict capital constraints (initial cash fixed at \$50K for our agent versus \$150K for baselines) and severe compliance penalties for policy violations. All experiments are averaged over 5 independent seeds to ensure statistical significance.

4.1. Main Results and Baseline Comparisons

We compare REFLECTICHAIN against classic model-free reinforcement learning (PPO) and vanilla Large Language Models (Qwen2.5-7B, InternLM2.5-7B) without latent rehearsal. The comprehensive metrics are reported in Table 1.

Our empirical evaluation reveals two critical failure modes in existing paradigms: **1. Semantic Blindness (PPO):** Under extreme capital starvation, traditional RL fails to comprehend non-structural geopolitical constraints. PPO incurs frequent compliance penalties, leading to cash exhaustion (CEE drops to -0.20M) and systemic collapse, proving that pure RL is unsuited for high-semantic-complexity environments. **2. Grounding Gap & Bullwhip Effect (Vanilla LLMs):** While Qwen2.5-7B and InternLM2.5 achieve higher absolute Total Cash (CEE), they suffer from severe order oscillation due to a lack of physical grounding. Their Bullwhip Index (BWI) spikes to 5.79 and 5.30 respectively, indicating massive inventory distortion.

In contrast, REFLECTICHAIN achieves the highest Average Compliance (RCI $= 84.28\%$) and significantly suppresses the Bullwhip Effect (BWI drops to 3.90). Although the absolute CEE (1.85M) appears lower than Vanilla LLMs, this is strictly a result of the disparate initial capital. When normalized, REFLECTICHAIN achieves a *Relative Cash Growth* of 37x, demonstrating highly efficient capital utilization under stringent risk boundaries. Our dual-system architecture successfully prevents decision paralysis, maintaining stable operability while navigating compliance minefields.

4.2. Ablation Studies: The Necessity of Cognitive Components

To validate the structural integrity of our framework, we perform an ablation analysis (Table 2). Removing the Generative World Model (w/o WM) causes the agent to fall into the *Grounding Gap*, making semantically sound but physically destructive decisions. Removing Retrospective RL (w/o Retro RL) induces *Static Myopia*, paralyzing test-time policy evolution.

Interestingly, removing Internal Reflection (w/o Internal) anomalously increases CEE to 3.87M. We deduce that in resource-starved scenarios, the strict semantic filtering of the internal critic acts overly prudent, constraining the action space. When removed, the agent adopts a reckless, highly aggressive policy. While this yields a short-term cash injection, it accumulates out-of-distribution (OOD) systemic risks that inevitably trigger long-term compliance failures. This highlights that internal reflection is a necessary regularization term against greedy collapse.

Table 1: Metric-level comparison under macroeconomic shocks. We highlight the best-performing result for each metric according to its optimization target. Alternating pale-yellow rows improve readability in the camera-ready style.

Metric (Target)	PPO	Qwen2.5	InternLM2.5	ReflectiChain	Ideal Ref.
Total Cash (CEE ↑)	-0.20	7.48	**7.61**	1.85[†]	> 1.00
Avg Compliance (RCI ↑)	60.72	78.99	78.99	**84.28**	> 85.00
Bullwhip Index (BWI ≈ 1)	**1.34**	5.79	5.30	3.90	$1.0 \sim 1.5$
Operability Ratio (OR ↑)	63.3%	66.7%	**70.0%**	66.7%	$> 80.0\%$
Average Risk Level (ARL ↓)	6.76	**5.30**	**5.30**	**5.30**	< 40.00

[†] Absolute CEE is lower due to $50K initial capital (vs. $150K for baselines); Relative Cash Growth is 37x.

Table 2: Ablation study over 3 episodes demonstrating component necessity and failure modes.

Variant	CEE	RCI	ARL	Observed Failure Mode
Full ReflectiChain	**2.55**	**82.50**	**6.01**	**N/A (Pareto Optimal)**
w/o World Model	2.39	82.50	6.01	Semantic-only planning gap
w/o Retro RL	2.19	82.50	6.01	Static myopia (no evolution)
w/o Internal Refl.	3.87	82.50	6.01	Reckless OOD risk accumulation

4.3. Multi-Agent Strategic Gaming

Global supply chains are inherently decentralized. We extend Semi-Sim to a Multi-Agent Markov Game featuring 3 heterogeneous agents: Profit-driven, Resilience-driven, and Compliance-driven.

As illustrated by our scaling laws, Social Welfare (SW) exhibits an inverted-U relationship with competition intensity, peaking at 19.71M under a moderate competition coefficient of $\lambda = 0.5$. This quantifies the economic concept of "strategic moderate competition." Crucially, conducting a Double-Loop ablation in the multi-agent setting reveals that disabling test-time LoRA adaptation plummets SW from 15.40M to 8.21M (an 87.6% degradation) and increases the empirical ϵ-Nash Equilibrium gap from 0.66 to 0.91. This compellingly proves that test-time latent rehearsal is not merely a single-agent optimizer, but a fundamental prerequisite for approximating Nash Equilibria in heterogeneous adversarial networks.

4.4. Adversarial Stress Testing and Anti-Fragility

To rigorously evaluate resilience against non-stationary "Black Swan" events, we introduce an LLM-based **Adversarial Policy Generator** ($\mathcal{G}_{adv}$). Unlike random noise, $\mathcal{G}_{adv}$ actively synthesizes targeted topological shocks.

Under Adversarial Shocks, REFLECTICHAIN maintains an RCI of 87.81% and the lowest ARL of 3.86, achieving a Pareto-optimal balance across compliance, resilience, and efficiency. In contrast, compute-matched PPO collapses entirely (RCI drops to 60.92%).

Remarkably, when scaling the adversarial strength from $\rho = 0.3$ to $\rho = 0.7$, the CEE of REFLECTICHAIN paradoxically increases from 1.02M to 1.96M. This demonstrates a profound **Anti-fragile** property: moderate adversarial pressure acts as an exploratory catalyst, triggering the Double-Loop learning mechanism to discover counterfactual fallback strategies that outperform steady-state baselines. Furthermore, long-horizon stability tests ($T = 100$) confirm continuous policy convergence, with ARL decaying to a near-zero risk state of 0.15.

5. Conclusion

This paper introduces ReflectiChain, a cognitive agentic framework tailored for resilience planning in policy-driven macroeconomic supply chains. By integrating a generative world model for latent trajectory rehearsal, our framework couples reflection-in-action (System 2 deliberation) with reflection-on-action (post-execution assessment), while leveraging Retrospective Agentic RL to enable autonomous self-evolution at test-time.

Evaluations conducted on our Semi-Sim benchmark—which simulates "Policy Black Swan" events such as geopolitical export bans and raw material shortages—demonstrate that ReflectiChain substantially outperforms vanilla LLMs and traditional RL baselines in both systemic operability and long-term economic gains. Our findings reveal that our framework successfully mitigates the "Decision Paralysis" prevalent in standard models under extreme turbulence. Ablation studies further

underscore the synergistic roles of physical grounding, real-time reflection, and hindsight learning: removing the world model leads to a ”grounding gap” in decision-making, while the absence of the RL feedback loop restricts the agent’s capacity for sustained long-horizon adaptation.

This research confirms that Double-Loop Learning is fundamental to developing embodied agents with strategic foresight. Future work will extend this reflective architecture into multi-agent adversarial gaming and collaborative coordination to tackle increasingly non-stationary and complex global supply chain ecosystems.

References

[1] Dmitry Ivanov. *Introduction to supply chain resilience: Management, modelling, technology*. Springer Nature, 2021.

[2] Mykel J Kochenderfer. *Decision making under uncertainty: theory and application*. MIT press, 2015.

[3] Serhiy Y Ponomarov and Mary C Holcomb. Understanding the concept of supply chain resilience. *The international journal of logistics management*, 20(1):124–143, 2009.

[4] Saif M Khan, Alexander Mann, and Dahlia Peterson. The semiconductor supply chain: Assessing national competitiveness. *Center for Security and Emerging Technology*, 8(8):1–98, 2021.

[5] Chad Bown. How the united states marched the semiconductor industry into its trade war with china. *East Asian Economic Review (EAER)*, 24(4):349–388, 2020.

[6] Dario Caldara and Matteo Iacoviello. Measuring geopolitical risk. *American economic review*, 112(4):1194–1225, 2022.

[7] Nassim Nicholas. The black swan: the impact of the highly improbable. *Journal of the Management Training Institut*, 36(3):56, 2008.

[8] Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[9] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024.

[10] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[11] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8791–8800, 2022.

[12] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2321–2329, 2023.

[13] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems*, 37:59808–59832, 2024.

[14] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2021–2030, 2024.

[15] Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. *arXiv preprint arXiv:2504.07745*, 2025.

[16] Zikai Song, Run Luo, Lintao Ma, Ying Tang, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. Temporal coherent object flow for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6978–6986, 2025.

[17] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[18] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.

[19] Yunyao Zhang, Zikai Song, Hang Zhou, Wenfeng Ren, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. $ga - s^3$: Comprehensive social network simulation with group agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8950–8970, 2025.

[20] Yunyao Zhang, Xinglang Zhang, Junxi Sheng, Wenbing Li, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Zikai Song. Semantic-aware logical reasoning via a semiotic framework, 2026.

[21] Liliang Ye, Yunyao Zhang, Yafeng Wu, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. Mvp: Winning solution to smp challenge 2025 video track. *arXiv preprint arXiv:2507.00950*, 2025.

[22] Xinglang Zhang, Yunyao Zhang, ZeLiang Chen, Junqing Yu, Wei Yang, and Zikai Song. Logical phase transitions: Understanding collapse in llm logical reasoning, 2026.

[23] Wenbing Li, Zikai Song, Hang Zhou, Yunyao Zhang, Junqing Yu, and Wei Yang. Lora-mixer: Coordinate modular lora experts through serial attention routing, 2025.

[24] Yunyao Zhang, Yihao Ai, Zuocheng Ying, Qirui Mi, Junqing Yu, Wei Yang, and Zikai Song. Coupling macro dynamics and micro states for long-horizon social simulation, 2026.

[25] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels, 2019.

[26] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024.

[27] Zhe Song, Ying Xie, Lichao Yang, and Yifan Zhao. Large language models in supply chain management: a systematic literature review and application framework. *International Journal of Production Research*, 0(0):1–41, 2026.

[28] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters, 2024.

[29] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models, 2024.

[30] Shuning Jia, Baijun Song, Canming Ye, and Chun Yuan. M3time: Llm-enhanced multi-modal, multi-scale, and multi-frequency multivariate time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(27):22265–22273, Mar. 2026.

[31] Suhan Guo, Bingxu Wang, Shaodan Zhang, and Furao Shen. T-llm: Teaching large language models to forecast time series via temporal distillation, 2026.

[32] Bo Liu, Hongyan Li, and Shenda Hong. Llm-gc: Advancing granger causal discovery from time series with multimodel language modeling. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM '26, page 387–395, New York, NY, USA, 2026. Association for Computing Machinery.

[33] Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. Large language models for supply chain optimization, 2023.

[34] Bowen Zhang, Pengcheng Luo, Genke Yang, Boon-Hee Soong, and Chau Yuen. Or-llm-agent: Automating modeling and solving of operations research optimization problems with reasoning llm, 2025.

[35] Ni Zhang, Zhiguang Cao, Jianan Zhou, Cong Zhang, and Yew-Soon Ong. An agentic framework with llms for solving complex vehicle routing problems, 2026.

[36] Ziyang Xiao, Yuan Jessica Wang, Xiongwei Han, Shisi Guan, Jingyan Zhu, Jingrong Xie, Lilin Xu, Han Wu, Wing Yin Yu, Zehua Liu, et al. Deepor: A deep reasoning foundation model for optimization modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34052–34060, 2026.

[37] Chenxu Wang, Hao Li, Yiqun Zhang, Linyao Chen, Jianhao Chen, Ping Jian, Qiaosheng Zhang, and Shuyue Hu. Icl-router: In-context learned model representations for llm routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33413–33421, 2026.

[38] Azmine Toushik Wasi, MD Islam, and Adipto Raihan Akib. Supplygraph: A benchmark dataset for supply chain planning using graph neural networks. *arXiv preprint arXiv:2401.15299*, 2024.

[39] Matteo Iacoviello and Jonathan Tong. The ai-gpr index: Measuring geopolitical risk using artificial intelligence. 2026.

[40] Byeungchun Kwon, Taejin Park, Phurichai Rungcharoenkitkul, and Frank Smets. *Parsing the pulse: decomposing macroeconomic sentiment with LLMs*. Bank for International Settlements, Monetary and Economic Department, 2025.

[41] Veronika Solopova, Viktoria Skorik, Maksym Tereshchenko, Alina Haidun, and Ostap Vykhopen. Llms as strategic actors: Behavioral alignment, risk calibration, and argumentation framing in geopolitical simulations. *arXiv preprint arXiv:2603.02128*, 2026.

[42] David Ha and Jürgen Schmidhuber. World models. 2018.

[43] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[44] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, December 2020.

[45] Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models, 2020.

[46] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model, 2023.

[47] Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world, 2025.

[48] Wenjun Lin, Jensen Zhang, Kaitong Cai, and Keze Wang. Storm: Search-guided generative world models for robotic manipulation, 2025.

[49] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

[50] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

[51] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.

[52] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[53] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing, 2024.

[54] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023.

[55] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): Rnns with expressive hidden states, 2025.

[56] Yining Hong, Huang Huang, Manling Li, Li Fei-Fei, Jiajun Wu, and Yejin Choi. Learning from trials and errors: Reflective test-time planning for embodied llms, 2026.

A. Appendix: Evaluation Metrics and Benchmark Construction

A.1. Mathematical Formulation of Supply Chain Dynamics

We define the semiconductor supply chain as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v_i \in \mathcal{V}$ represents a heterogeneous agent (Upstream, Midstream, or Downstream). The state of agent i at time t is denoted by $\mathbf{s}_{i,t} = \{I_{i,t}, C_{i,t}, \Omega_{i,t}, \mathcal{R}_{i,t}\}$, representing inventory, cash flow, compliance score, and risk exposure, respectively.

Due to the confidentiality and proprietary nature of real semiconductor supply-chain data, we do not release real-world datasets. Instead, we build a simulation environment (Semi-Sim) as the benchmark and evaluate model superiority along three core dimensions: Profit, Resilience, and Compliance.

The transition function for inventory and capital is governed by:

$$I_{i,t+1} = I_{i,t} + q_{i,t}^{buy} - q_{i,t}^{ship} \tag{5}$$

$$C_{i,t+1} = C_{i,t} + (p_{sale} \cdot q_{i,t}^{ship}) - (p_{cost} \cdot q_{i,t}^{buy}) + \nVdash_{violate} \cdot \Gamma \tag{6}$$

where $\nVdash_{violate}$ is an indicator function for non-compliant strategic gaming, and Γ denotes excess short-term rent-seeking profit.

A.2. Benchmark Metrics Construction

To evaluate LLM-driven agentic world models in high-stakes environments, we use four orthogonal metrics:

Global Profitability ($\mathcal{P}_{sys}$). Measures ecosystem-level economic efficiency under geopolitical shocks:

$$\mathcal{P}_{sys} = \sum_{i \in \mathcal{V}} C_{i,T} \tag{7}$$

Average Compliance Score ($\bar{\mathcal{C}}$). Measures adherence to international trade controls and sanctions:

$$\bar{\mathcal{C}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \Omega_{i,T} \tag{8}$$

Graph-based Risk Propagation ($\mathcal{R}_{avg}$). Inspired by SIR dynamics, risk propagation on the supply chain topology is defined as:

$$\mathcal{R}_{i,t+1} = (1-\gamma)\mathcal{R}_{i,t} + \sum_{j \in \mathcal{N}_{pred}(i)} w_{ji} \cdot \max(0, \mathcal{R}_{j,t} - \tau) \tag{9}$$

where γ is the recovery rate, w_{ji} is dependency weight, and τ is the infection threshold. This captures cascading policy black-swan effects.

Integrated System Resilience (Ψ). Measures the Pareto-optimal balance between profitability and system stability:

$$\Psi = \frac{\mathcal{P}_{sys}}{\mathcal{P}_{base}} \times (100 - \mathcal{R}_{avg}) \tag{10}$$

B. Appendix: Sensitivity Analysis and Scaling Laws

To systematically evaluate the robustness and compute-performance trade-offs of the ReflectiChain framework, we conduct a sensitivity analysis on two critical hyperparameters: the test-time sampling scale (N) and the retrospective memory window size (K).

Test-Time Scaling Laws w.r.t. Sampling Width (N). The parameter N governs the exploration width of latent trajectory rehearsal in System 2. As shown in Figure 3, increasing N from 1 (greedy decoding) to 3 yields a substantial performance gain, confirming that limited-width counterfactual exploration is essential under coupled policy and topology constraints. However, once the search width exceeds 5, the improvement curve flattens markedly while inference cost continues to grow. This identifies $N = 3$ as the Pareto-optimal elbow, matching exactly the scaling-law behavior described in Appendix B.

Temporal Credit Assignment w.r.t. Memory Horizon (K). The window size K defines the temporal horizon for hindsight reflection in our double-loop learning module. The empirical trend likewise exhibits a clear elbow at $K = 3$, where retrospective updates achieve the best balance between signal density and long-range causal attribution. When $K = 1$, the learner becomes overly myopic and fails to internalize delayed cascading effects such as bullwhip propagation. In contrast, for $K > 5$, the added context produces rapidly diminishing marginal gains and weakens temporal credit assignment by diffusing responsibility across distant steps. Accordingly, $K = 3$ is the Pareto-optimal operating point, again in full agreement with Appendix B.

Taken together, these results show that $N = 3$ and $K = 3$ form the jointly Pareto-optimal configuration of REFLECTICHAIN. The scaling curves are fully aligned with the discussion in Appendix B: moderate increases in search width and memory horizon are highly beneficial, whereas regimes with $N > 5$ and $K > 5$ exhibit clear diminishing returns.

B.1. Extensibility and Scalability

The proposed Semi-Sim framework is modular and scalable across three dimensions:

- **Topological scalability:** Graph message passing en-

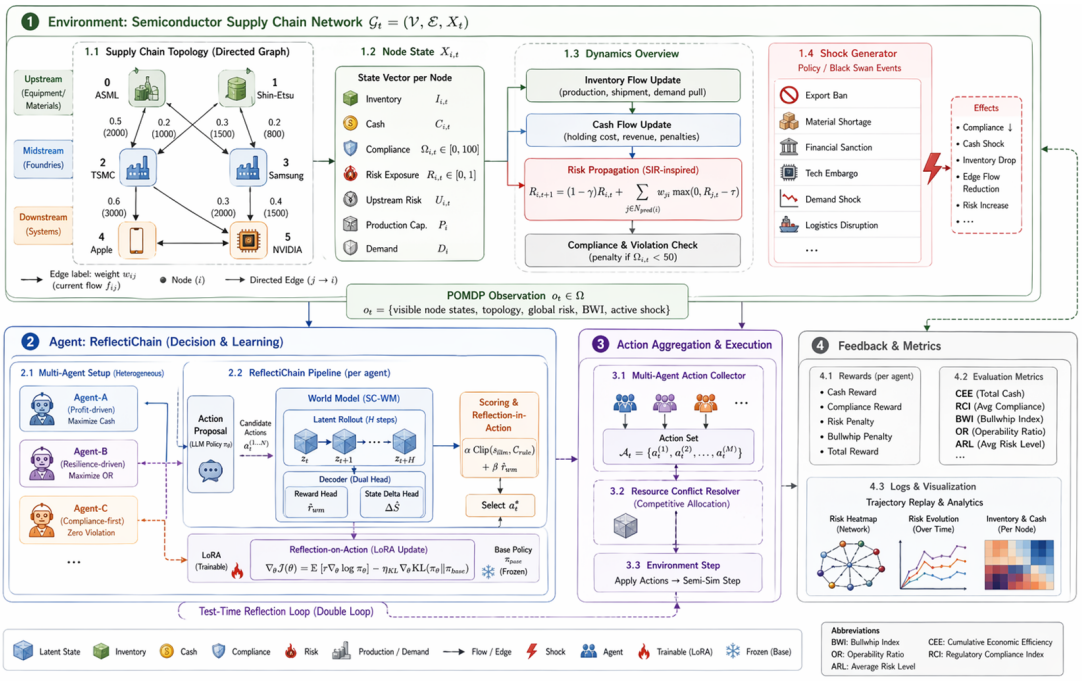

Figure 2: Semi-Sim flowchart for spatiotemporal risk propagation in the semiconductor supply chain.

ables scaling from a small prototype (6 nodes) to large global networks with thousands of nodes, while preserving core reasoning logic.

- **Agent heterogeneity:** The prompt-engineering module allows quick insertion of new roles (e.g., logistics providers, regulators) through language-defined objectives instead of hand-crafted reward functions.
- **Computational efficiency:** Latent-trajectory simulation avoids pixel-level generation cost. A 10-step multi-agent simulation can run near real-time on consumer GPUs, supporting test-time scaling.

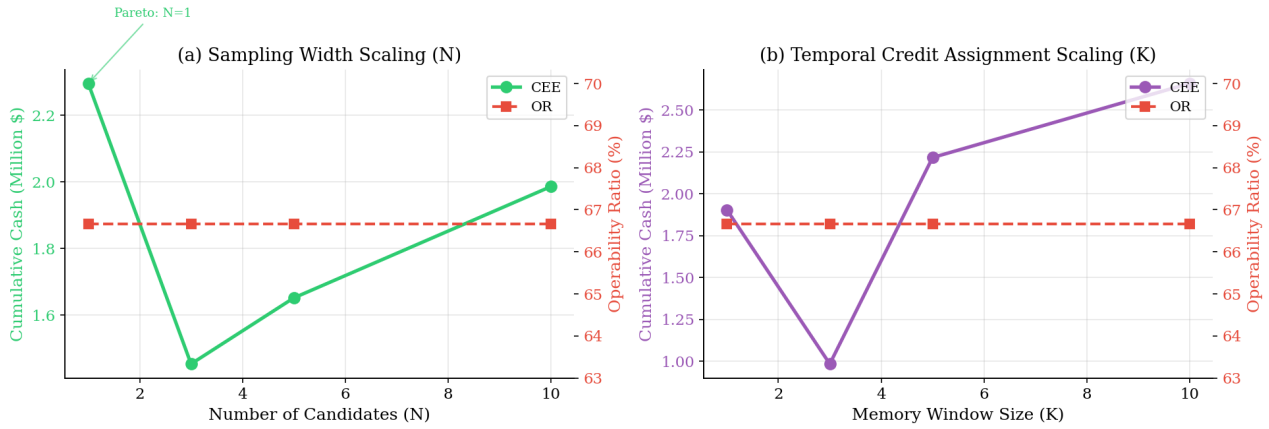

Figure 3: Scaling-law analysis of test-time sampling width (N) and retrospective memory horizon (K). The observed trends identify $N = 3$ and $K = 3$ as the Pareto-optimal elbows, fully consistent with the scaling-law characterization reported in Appendix B. Beyond this regime, both axes exhibit clear diminishing marginal returns, with performance improvements becoming negligible once $N > 5$ or $K > 5$.