Available Online: 14 January 2026 DOI: 10.54254/2977-5701/2026.31322

A risk-constrained distributional reinforcement learning portfolio strategy integrating chain-of-thought logical priors and time-varying topological structures

Yuting Wang

Renmin University of China, Beijing, China

bay.w@outlook.com

Abstract. Time series of financial asset returns typically exhibit pronounced non-ergodicity and heavy-tailed, spiky distributions. Traditional mean–variance models and expectation-based Deep Reinforcement Learning (DRL) approaches struggle to effectively capture distributional shifts induced by exogenous logical shocks. To address this challenge, this paper proposes a risk-constrained distributional reinforcement learning framework that integrates large language model–based logical reasoning with dynamic graph dependencies (LLM-G-DRL). At the perception level, rather than relying on conventional sentiment polarity classification, this study introduces large language models that support Chain-of-Thought (CoT) reasoning (e.g., DeepSeek-R1) to construct a Bayesian logical belief updating mechanism. This mechanism maps unstructured financial texts into high-dimensional latent logical states embedding causal transmission paths, thereby correcting predictive biases arising from exclusive dependence on historical price and volume data. At the structural level, to characterize the nonlinear contagion of systemic risk, the framework abandons the assumption of static adjacency matrices and employs Dynamic Graph Attention Networks (Dynamic GAT) to reconstruct time-varying topological dependencies among assets, enabling explicit modeling of risk propagation channels. At the decision-making level, the portfolio optimization problem is formulated as a Constrained Markov Decision Process (CMDP). Implicit Quantile Networks (IQN) are adopted to approximate the full probability distribution function while preserving higher-order moment information. Furthermore, based on Lagrangian duality theory, a hard constraint on Conditional Value at Risk (CVaR) is introduced, transforming tail-risk control into a dynamic penalty mechanism governed by dual variables. Theoretical analysis demonstrates that, through a closed-loop design combining "logical priors, structural contagion, and distributional decision-making," the proposed framework exhibits stronger mathematical robustness than traditional point-estimation models. It effectively identifies and avoids tail losses under extreme market conditions, offering a statistically interpretable new paradigm for intelligent asset allocation in non-stationary markets.

Keywords: distributional reinforcement learning, chain-of-thought logical priors, time-varying topological dependence, Conditional Value at Risk (CVaR), constrained Markov decision process

1. Introduction

1.1. Research background: a paradigm shift from data mining to logical inference

The evolution of financial markets is essentially a high-dimensional, non-stationary stochastic dynamical process characterized by substantial noise. Within the classical econometric framework, asset allocation theories represented by the mean–variance model of Markowitz typically rely on strong assumptions, including joint normality of asset returns and time-invariant correlation structures [1]. However, as noted by Yongmiao Hong and Shouyang Wang, financial markets in the era of big data exhibit pronounced non-ergodic behavior, rendering traditional low-dimensional statistics insufficient for capturing distributional shifts induced by exogenous shocks such as macroeconomic policy shifts or sudden geopolitical events [2].

Confronted with this challenge, the focus of quantitative research has gradually shifted from the mere mining of structured price–volume data toward the interpretation of massive volumes of unstructured information. Although early financial text-mining studies (e.g., FinBERT) achieved notable success in sentiment classification tasks through pre-trained language models [3], their methodological essence remains confined to supervised "sentiment polarity classification," lacking the capacity to infer the complex causal chains underlying financial events. Recently, with the emergence of Reinforcement Learning–incentivized Large Language Models (RL-Incentivized LLMs) such as DeepSeek-R1, machine intelligence has, for the first time, demonstrated the ability to perform multi-step reasoning via Chain-of-Thought (CoT) mechanisms [4, 5]. This development provides a fundamentally new statistical modeling paradigm for transforming unstructured text into financially meaningful "logical prior factors," thereby correcting the overfitting tendencies inherent in purely data-driven models.

1.2. Limitations of existing research and theoretical gaps

Despite the widespread application of Deep Reinforcement Learning (DRL) in algorithmic trading and portfolio management due to its strong performance in sequential decision-making problems [6, 7], prevailing frameworks continue to face three major theoretical gaps when dealing with complex financial systems.

First, there is a "semantic deficiency" in information states. Most existing DRL agents merely concatenate textual sentiment scores into the state space, overlooking the heterogeneity of market logic. For example, an interest rate cut propagates through highly leveraged industries and export-oriented sectors via fundamentally different transmission pathways; such structured differences cannot be captured by simplistic embeddings that lack logical reasoning.

Second, there is a "static topology" assumption in risk transmission. Systemic risk often exhibits pronounced network contagion effects [8]. Nevertheless, the majority of studies still rely on static Pearson correlation matrices or fixed industry classifications to define asset relationships. Empirical evidence provided by Wang et al. indicates that inter-asset dependencies evolve dynamically [9], and static graph structures fail to capture the surge in correlations during crises that gives rise to tail dependence.

Finally, there exists a "risk-neutral bias" in decision objectives. Traditional algorithms such as DDPG or PPO aim to maximize the expected cumulative return, which is statistically equivalent to assuming risk-neutral investors. Although distributional reinforcement learning, as proposed by Bellemare et al., demonstrates that learning the full value distribution is superior to learning a single expected value, how to integrate this framework with the convex optimization theory of Conditional Value at Risk (CVaR) introduced by Rockafellar, and thereby impose hard tail-risk constraints in an end-to-end learning process, remains an unresolved mathematical challenge.

1.3. Research agenda and contributions of this study

To address the aforementioned challenges, this paper proposes a risk-constrained distributional reinforcement learning framework that integrates unstructured logical priors with dynamic graph dependencies (LLM-G-DRL). The central idea of this study is to abandon the traditional "prediction–optimization" separation paradigm and instead construct a closed-loop system that unifies "logical perception (DeepSeek), structural transmission (Dynamic Graph), and distributional decision-making (IQN)."

The main theoretical contributions of this paper are threefold:

1. Construction of a Chain-of-Thought–Based Bayesian Logical State Space: Unlike conventional semantic vectorization approaches, this study leverages the reasoning capabilities of DeepSeek-R1 to extract "event–pathway–impact" logical chains from financial news and map them into high-dimensional latent state vectors. From a statistical perspective, this is equivalent to introducing an exogenous logical prior, which effectively reduces estimation variance in small-sample, high-noise environments.

2. A Logic-Driven Time-Varying Topology Learning Mechanism: Building on the ideas of MRRFGNN [9], this paper develops a dynamic graph attention network that allows inter-asset adjacency relationships to adaptively evolve with market logical states, thereby explicitly modeling the nonlinear contagion of systemic risk within a portfolio.

3. A Distributional Decision Model with CVaR Dual Constraints: From the perspective of stochastic optimal control, portfolio optimization is formulated as a Constrained Markov Decision Process (CMDP). Implicit Quantile Networks (IQN) are employed to approximate the full probability distribution of returns, while Lagrangian relaxation techniques are used to embed CVaR tail-risk constraints into a dual gradient descent algorithm, enabling joint optimization of the return distribution and risk boundaries.

2. Theoretical foundation and problem formulation

This chapter first formalizes the multi-source information–driven portfolio management problem as a Constrained Markov Decision Process (CMDP). It then elaborates on the modeling of inter-asset dependency structures based on dynamic graph attention networks, and finally derives the dual representations of the value function in distributional reinforcement learning and the CVaR risk measure grounded in quantile regression theory.

2.1. Problem definition: Constrained Markov Decision Process (CMDP)

Portfolio management is essentially a sequential decision-making problem under an uncertain stochastic environment. Given institutional investors' stringent requirements on tail-risk control, we formulate the problem as a five-tuple $\mathscr{M} = \langle S, A, P, \mathscr{R}, C \rangle$.

1. State Space S At time t

$s_t \in S$ is represented as a multimodal feature set encompassing three heterogeneous information streams:

$$s_t = \{X_t^{quant}, G_t, h_t^{logic}\} \tag{1}$$

- Micro-level price and volume features $X_t^{quant} \in \mathrm{R}^{N \times F \times T_w}$: Contains Ntarget assets' open price, close price, trading volume, and F -dimensional technical indicators over a lookback window of T_w .

- Dynamic topological structure G_t : A time-varying adjacency matrix of asset relationships extracted via a graph neural network (see Section 2.2).

- Logical latent state $h_t^{logic} \in \mathrm{R}^{N \times D_{LLM}}$: High-dimensional causal-logic vectors derived from unstructured financial text streams using the DeepSeek-R1 model's Chain-of-Thought (CoT) reasoning capability [4, 5].

These vectors encode market sentiment while implicitly capturing the expected transmission pathways of events to asset prices.

2. Action Space A

$a_t \in A \subseteq \mathrm{R}^{N+1}$ represents the portfolio weight vector at time t. Let $a_{t,0}$ denote the weight of the risk-free asset (cash) and at $a_{t,i}$ denote the weight of the i-th risky asset. To satisfy practical financial constraints, the action vector is subject to non-negativity and simplex normalization:

$$\sum_{i=0}^{N} a_{t,i} = 1, a_{t,i} \geq 0, \forall i \in \{0, \ldots, N\} \tag{2}$$

3. Reward Function $\mathscr{R}$

The logarithmic portfolio return at time t is defined as:

$$r_t(s_t, a_t) = ln(\sum_{i=0}^{N} a_{t,i} \cdot \frac{p_{t+1,i}}{p_{t,i}} \cdot (1 - \delta_i)) \tag{3}$$

where $p_{t,i}$ denotes the price of asset i at time t, and δ_i represents the transaction cost rate including slippage and fees.

4. Constraints C

To control tail losses under extreme market conditions, we introduce a hard CVaR (Conditional Value at Risk) constraint. The policy π is required to ensure that the downside risk at any time does not exceed a predefined threshold ξ:

$$J_C(\pi) = CVaR_{\alpha}[-r_t(s_t, a_t)] \leq \xi \tag{4}$$

2.2. Dynamic graph modeling of asset dependencies

Financial asset dependencies exhibit pronounced time-varying and nonlinear characteristics. Studies indicate that during periods of market stress, systemic risk propagates via cascading effects across financial networks [8]. To capture such evolving dependencies, we construct a dynamic graph: $G_t = (V, \mathscr{E}_t)$.

Node representation: The node set $V = \{v_1, \ldots, v_N\}$ represents all assets in the market. Each node's initial embedding integrates micro-level price-volume features with LLM-derived textual logic features:

$$e_{i,t} = \phi_{fusion}(x_{i,t}^{quant}, h_{i,t}^{logic}) \tag{5}$$

Dynamic edge generation: Unlike conventional static adjacency matrices based on Pearson correlation, we adopt a multi-relation reconstruction approach inspired by Wang et al. [9], using a Multi-Head Graph Attention Network (GAT) [10] to dynamically learn attention coefficients $\alpha_{ij,t}$ between assets i and j:

$$\alpha_{ij,t} = \frac{exp(LeakyReLU(a^T[We_{i,t} \| We_{j,t}]))}{\sum_{k \in N_i} exp(LeakyReLU(a^T[We_{i,t} \| We_{k,t}]))} \tag{6}$$

Here, $||$ denotes vector concatenation, and W and$a$ are learnable parameters. This mechanism allows the model to adaptively adjust contagion weights according to market conditions (e.g., during crises).

2.3. Distributional reinforcement learning and Implicit Quantile Networks (IQN)

Traditional Reinforcement Learning (RL) maximizes expected returns $E[Q(s, a)]$, neglecting higher-order moments of the return distribution (e.g., skewness, kurtosis). In this work, we adopt a Distributional RL framework [11], modeling the cumulative return as a random variable $Z^{\pi}(s, a)$:

$$Z^{\pi}(s, a)r(s, a) + \gamma Z^{\pi}(s', a') \tag{7}$$

where $\overset{D}{=}$ denotes distributional equivalence.

To approximate this distribution, we employ an Implicit Quantile Network (IQN) [12], which learns a mapping from quantile $\tau \sim U([0,1])$ to value, $F_Z^{-1}(\tau)$, instead of discretizing the distribution. For a given state-action pair (s,a) and quantile τ, IQN outputs the corresponding quantile estimate $Z_\tau(s,a)$.

Quantile embedding: To allow the network to capture the position of τ, a cosine basis function maps the scalar τ into a high-dimensional vector:

$$\phi(\tau) = ReLU(\textstyle\sum_{k=0}^{n-1} cos(k\pi\tau)w_k + b_k) \tag{8}$$

Distributional loss function: For two quantile samples τ, τ', the Huber quantile regression loss updates network parameters:

$$\mathscr{L}_{IQN}(\theta) = E_{\tau,\tau'\sim U([0,1])}[\rho^{\kappa}_{\tau}(r + \gamma Z_{\tau'}(s', \pi(s')) - Z_\tau(s,a))] \tag{9}$$

This allows the model to reconstruct the full probability density function of returns, effectively capturing "fat tails" and extreme events in financial markets.

2.4. CVaR-based risk-constrained dual formulation

To pursue high returns while controlling extreme risk, we introduce Conditional Value at Risk (CVaR) into the optimization objective. Following Rockafellar & Uryasev [13], the $CVaR_\alpha$ at confidence level α is defined as the conditional expectation over the worst $(1-\alpha)$ tail of the loss distribution.

Within the IQN framework, since the quantile function $Z_\tau(s,a)$ is directly accessible, CVaR can be efficiently computed via discrete integration over the left-tail quantiles:

$$\widehat{CVaR}_\alpha(s,a) = -\tfrac{1}{K}\textstyle\sum_{k=1}^{K} Z_{\tau_k}(s,a), \tau_k \sim U([0,\alpha]) \tag{10}$$

To handle the constrained optimization problem, we leverage Lagrangian duality. Introducing a non-negative Lagrange multiplier $\lambda \geq 0$, the original constrained problem is reformulated as an unconstrained min-max problem:

$$\mathscr{L}_{total}(\theta, \lambda) = -E[Z^\pi] + \lambda \cdot softplus(\widehat{CVaR}_\alpha(Z^\pi) - \xi) \tag{11}$$

Here, λ acts as a risk penalty factor. When the predicted CVaR exceeds the threshold ξ, λ increases automatically, guiding the policy gradient toward risk-reducing directions, achieving a dynamic balance between return and risk.

3. Proposed decision-making framework: LLM-G-DRL

This chapter formally introduces an intelligent portfolio decision-making framework that integrates large language model priors with graph-based distributional reinforcement learning—namely, the Large Language Model–based Graph Distributional Reinforcement Learning (LLM-G-DRL) framework. Addressing three core challenges of financial markets—high-dimensional nonlinearity, time-varying dependency structures, and the latent nature of tail risks—the proposed framework abandons the traditional "prediction–optimization" separation paradigm and instead constructs an end-to-end closed-loop system encompassing perception, cognition, and decision-making.

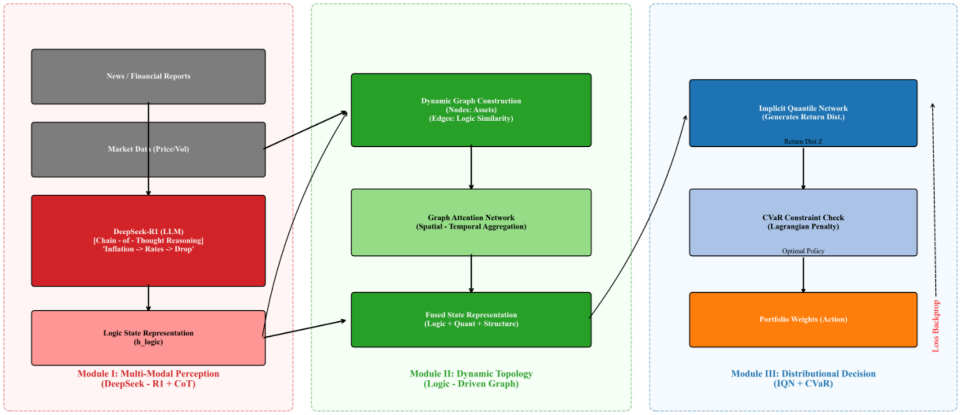
News / Financial Reports
Market Dat (Price/Vol)
DeepSeek-R1 (LLM)
[Chain - of - Thought Reasoning]
'Inflation -> Rates -> Drop'
Logic State Representation
(h_logic)
Module I: Multi-Modal Perception
(DeepSeek - R1 + CoT)
Dynamic Graph Construction
(Nodes: Assets)
(Edges: Logic Similarity)
Graph Attention Network
(Spatial - Temporal Aggregation)
Fused State Representation
(Logic + Quant + Structure)
Module II: Dynamic Topology
(Logic - Driven Graph)
Implicit Quantile Network
(Generates Return Dist.)
Return Dist Z
CVaR Constraint Check
(Lagrangian Penalty)
Optimal Policy
Portfolio Weights (Action)
Loss Backprop
Module III: Distributional Decision
(IQN + CVaR)

Figure 1. The overall architecture of the LLM-G-DRL framework

As illustrated in Figure 1, the overall architecture consists of three tightly coupled submodules: a multimodal state encoder based on Bayesian semantic fusion, a logic-driven dynamic graph dependency learner, and a risk-constrained distributional policy network.

3.1. Module I: multimodal state encoding based on Bayesian semantic fusion

The financial market state s_t is determined not only by realized price–volume data (posterior observations), but also by market expectations regarding future events (prior beliefs). This module reconstructs the information state space through a dual-stream architecture.

1. Structured Feature Extraction (Micro-Level Price–Volume Stream): For the historical window data ($X_{i,t}^{price} \in \mathrm{R}^{T\times F}$) of asset i at time t, a one-dimensional causal convolutional neural network (Causal 1D-CNN) is employed to extract local temporal feature representations $h_{i,t}^{quant}$. Through a masking mechanism, causal convolutions ensure that convolutional kernels only cover inputs prior to the current time step, thereby strictly preventing look-ahead bias and information leakage.

2. Unstructured Logical Reasoning (Macro-Level Semantic Stream): This component constitutes one of the core innovations of the proposed framework. Unlike conventional BERT-based sentiment classification approaches, we introduce the DeepSeek-R1 large language model as a logical reasoning engine.

• Chain-of-Thought (CoT) Prompting: We design structured prompts incorporating "event–transmission pathway–expected impact" elements to guide the LLM toward generating explicit reasoning traces.

• Semantic Vectorization: The reasoning text produced by the LLM is fed into a pre-trained Transformer encoder to obtain a latent semantic vector $h_{i,t}^{text}$. This vector encodes not only sentiment polarity, but also embedded causal logic such as "interest rate hikes → liquidity tightening → valuation compression of technology stocks."

3. Bayesian Gated Fusion: To address the dynamic variation in the relative importance of different information sources across market regimes (e.g., price–volume dominance during tranquil periods versus news dominance during crises), an adaptive gating mechanism is designed to generate the final node representation $x_{i,t}$:

$$z_t = \sigma(W_z[h_{i,t}^{quant} \| h_{i,t}^{text}] + b_z) \tag{12}$$

$$x_{i,t} = z_t \odot h_{i,t}^{quant} + (1 - z_t) \odot (W_p h_{i,t}^{text}) \tag{13}$$

where $\odot$ denotes the Hadamard product. The gating variable z_t functions as a confidence weight, mimicking a Bayesian belief-updating process.

3.2. Module II: logic-driven time-varying graph dependency learning

To faithfully replicate historical trading environments, the proposed framework strictly adheres to the Point-In-Time (PIT) principle. In response to shifts in inter-asset correlation structures driven by evolving market logic —such as sector rotation or supply chain disruptions—this module constructs dynamic graph structures inspired by the MRRFGNN framework [9].

1. Dynamic Graph Topology Inference

The asset correlation graph at time t is defined as: $G_t = (V, \mathcal{E}_t)$. To explicitly model the effect of logical factors on asset correlations, we incorporate interactions of logical features into the attention coefficients:

$$e_{ij,t} = LeakyReLU(a^T[Wx_{i,t} \| Wx_{j,t} \| \beta \cdot sim(h_{i,t}^{text}, h_{j,t}^{text})]) \tag{14}$$

Here, $sim(\cdot)$ denotes cosine similarity, and β is a logic influence factor. This design allows the graph network to rapidly establish connections when two stocks face similar macro-logical shocks (e.g., both impacted by rising oil prices), even if their historical price-volume correlation is low, enabling early warning of risk contagion.

2. Logic-Enhanced Graph Feature Aggregation

Using the normalized attention weights $\alpha_{ij,t}$, neighbor node information $\widetilde{x}_{i,t}$ is aggregated to obtain a market-contextualized asset representation:

$$\widetilde{x}_{i,t} = \sigma(\textstyle\sum_{j \in N_i} \alpha_{ij,t} W x_{j,t}) \tag{15}$$

3.3. Module III: risk-constrained distributional actor

To effectively handle the fat-tailed nature of financial data, this framework abandons traditional DRL that only predicts expected returns, instead adopting distributional reinforcement learning with CVaR hard constraints.

1. IQN-Based Return Distribution Modeling: The cumulative return of the portfolio is modeled as a random variable $Z^\pi(s,a)$. The IQN network takes the state s_t(aggregated from $\widetilde{x}_t$) and a random quantile sample $\tau \sim U([0,1])$ as input, outputting the corresponding quantile return $Z_\tau(s_t, a_t)$. This internally simulates the full probability distribution of future returns, allowing the agent to know not only "how much it can earn on average" but also "how much it might lose in the worst case."

2. CVaR-Constrained Lagrangian Relaxation: Based on the dual formulation in Section 2.4, the total loss function is defined as: $L(\theta, \lambda)$. This function comprises two components:

- Quantile Regression Loss: accurately fits the return distribution Z^π.
- CVaR Penalty (Risk Penalty): $\lambda \cdot softplus(CVaR_\alpha(Z^\pi) - \xi)$.

During training, a primal-dual update strategy is adopted:

- Policy network parameters θ update: minimize the total loss to find a policy that maximizes return while satisfying risk constraints.
- Lagrange multiplier λ update: maximize the total loss (gradient ascent). When the risk constraint is violated, λ\lambda$\lambda$ increases, forcing the policy back into a risk-safe region.

3.4. Algorithm workflow

To clearly illustrate the training logic of the proposed framework, the overall training procedure of LLM-G-DRL is summarized in Algorithm 1.

Algorithm 1: LLM-G-DRL Training with Risk Constraints

```
Input: Multimodal data stream, DeepSeek-R1 model, Risk threshold $\xi$, Confidence level $\alpha$ .
Initialize: Policy network parameters $\theta$, Dual variable $\lambda$ .
For each episode do:
    For each time step $t$ do:
        Cognitive Step: Extract logic vector $h_t^{logic}$ via DeepSeek-R1 CoT.
        Structural Step: Construct dynamic graph $G_t$ and aggregate features to get $s_t$.
        Decision Step: Sample $K$ quantiles $\tau_k \sim U([0,1])$, compute action $a_t = argmax_a \frac{1}{K} \sum Z_{\tau_k}(s_t, a)$ .
        Execute $a_t$, observe reward $r_t$ and next state $s_{t+1}$ . Store transition in Replay Buffer.
        Optimization Step (Mini-batch):
            Sample batch transitions.
            Compute Quantile Loss $\mathscr{L}_{IQN}$ and CVaR constraint violation.
            Update $\theta \leftarrow \theta - \eta_1 \nabla_\theta(\mathscr{L}_{IQN} + \lambda \cdot Penalty)$.
            Update $\lambda \leftarrow \lambda + \eta_2 \nabla_\lambda(\lambda \cdot Penalty)$.
    End For
End For
```

4. Theoretical analysis and mechanism discussion

This chapter does not aim to present overfitted performance curves from a single historical backtesting window. Instead, from statistical and topological perspectives, it provides an in-depth analysis of the intrinsic mechanisms through which the LLM-G-DRL framework addresses non-stationary financial distributions, systemic risk contagion, and extreme tail events.

4.1. Robustness of logical priors against distributional shift

Traditional quantitative models (e.g., LSTM, Transformer) are essentially based on Empirical Risk Minimization (ERM) over historical data D_{hist} . According to statistical learning theory, their effectiveness relies on the independent and identically distributed (i.i.d.) assumption. However, financial markets exhibit pronounced distributional shifts, namely, $P_{train}(X,Y) \neq P_{test}(X,Y)$.

Proposition 1: Introducing the logical state h_{logic} generated by a large language model (DeepSeek-R1) is equivalent to incorporating a time-varying Bayesian prior, which effectively reduces the model's generalization error in non-stationary environments.

Analysis: In purely data-driven models, when confronted with unprecedented "black swan" events (e.g., the circuit breakers in 2020), the quantitative feature vector x_{quant} falls outside the manifold spanned by the training data, resulting in a sharp divergence of the prediction variance $Var(\hat{y})$. In contrast, within the LLM-G-DRL framework, we introduce a latent logical variable $z$generated via LLM-based reasoning. Even when price–volume patterns appear unfamiliar, the underlying causal logic (e.g., "liquidity depletion → asset sell-offs") remains dense and generalizable in semantic space. Mathematically, this amounts to constructing a conditional predictive model $P(Y \mid X, Z)$. Since Z (logical information) exhibits cross-regime invariance as an underlying mechanism, its inclusion significantly constrains the model's search space in Out-Of-Distribution (OOD) regions. As a result, the framework achieves theoretically stronger robustness than purely end-to-end data-driven models.

4.2. Systemic risk interruption via dynamic topology

Existing studies [8, 9] indicate that systemic risk often manifests as an abrupt breakdown of correlation structures, wherein inter-asset correlations rapidly converge toward unity. Static graph neural networks are inherently incapable of capturing such topological phase transitions.

Mechanism Analysis: The dynamic graph attention mechanism in the proposed framework embeds a two-layer defense logic:

1. Logical Resonance Identification: The attention coefficient α_{ij} explicitly depends on the logical similarity $sim(h_i^{text}, h_j^{text})$. When DeepSeek identifies that assets within the "semiconductor sector" are exposed to a shared logical shock (e.g., "export restrictions"), the graph network instantaneously establishes strong connections (high attention weights) between them, even if historical correlations were weak.

2. Risk Contagion Interruption: Under the CVaR constraint, when the risk estimate Z_τ of a particular node (e.g., a sector leader) deteriorates, these high-weight connections rapidly propagate negative signals to neighboring nodes, triggering defensive rebalancing actions downstream. This mechanism emulates the default cascade defense described in the Gai–Kapadia model, effectively severing contagion chains before systemic risk fully materializes.

4.3. Geometric interpretation of tail risk control

Traditional mean–variance optimization attempts to compress return distributions into a compact Gaussian sphere, an assumption that frequently breaks down under skewed and heavy-tailed market conditions. By contrast, the integration of IQN and CVaR constraints introduces a novel mechanism of distributional geometric clipping.

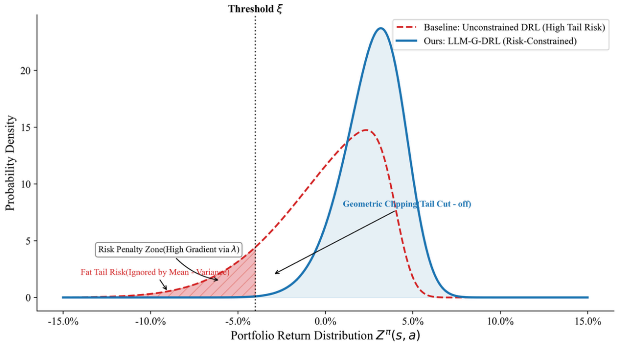

Figure 2. Geometric interpretation of CVaR-based distributional clipping

As illustrated in Figure 2:

- Unconstrained DRL (e.g., DQN, PPO): tends to aggressively stretch the right tail of the return distribution in pursuit of higher expected gains, often at the cost of an excessively elongated left tail that exposes the portfolio to catastrophic losses.
- LLM-G-DRL: through adaptive adjustment of the Lagrange multiplier λ, alters the gradient direction once the integral over the predicted left-tail quantiles (i.e., CVaR) exceeds the threshold ξ . Geometrically, this manifests as a hard truncation of the left side of the probability density function. The model is thus compelled to abandon actions with attractive expected returns but excessive downside exposure (e.g., highly leveraged long positions), in favor of strategies with more compact and defensively shaped return distributions.

4.4. Qualitative case study: strategy evolution under anticipated federal reserve rate hikes

To illustrate the perception–reasoning–decision closed loop of the proposed framework, we construct a thought experiment based on an unexpected Federal Reserve rate hike.

Stage 1: Logical Perception (DeepSeek-R1)

• Input: Financial news streams report "CPI data exceeds expectations; the Federal Reserve adopts a hawkish stance."

• Reasoning: DeepSeek generates the following chain of logic: Persistently high inflation → rising probability of rate hikes → higher risk-free rates → larger discount factors in equity valuation models → downward pressure on growth stock prices.

• Output: A negative logic vector h_{tech}^{logic} targeting the technology sector, and a positive logic vector h_{bank}^{logic} favoring banking stocks that benefit from widening interest margins.

Stage 2: Structural Propagation (Dynamic Graph Network)

• The graph network detects a surge in logical similarity among technology stocks and automatically strengthens intra-sector edge weights. Consequently, a minor price decline in a leading tech stock is amplified through the graph structure and rapidly propagated across the entire sector.

Stage 3: Distributional Decision-Making and Risk Control (IQN + CVaR)

• The IQN predicts a pronounced left-skewed return distribution for technology stocks, causing the estimated CVaR to breach the predefined threshold ξ.

• Dual response: The Lagrange multiplier λ increases sharply.

• Final action: To minimize the overall loss, the policy substantially reduces exposure to technology equities while reallocating capital toward cash holdings and low-valuation defensive sectors.

This qualitative analysis demonstrates that LLM-G-DRL is not a purely black-box fitting procedure. Instead, it embodies an interpretable macro-hedging logic grounded in economic reasoning and tail-risk control—capabilities that are largely absent from traditional quantitative models.

5. Conclusion and future directions

5.1. Summary of findings

Addressing the high-dimensional non-stationarity and systemic risk contagion inherent in financial markets, this study proposes an integrated portfolio decision-making framework—LLM-G-DRL—that fuses large language model–based logical reasoning with dynamic graph–based distributional reinforcement learning. Rather than relying on naïve multimodal feature concatenation, the framework reconstitutes the mathematical foundations of quantitative investment from three complementary dimensions: perception mechanisms, structural modeling, and decision paradigms.

1. From sentiment polarity to logical causality in perception: By incorporating the Chain-of-Thought (CoT) reasoning capability of DeepSeek-R1, the processing of unstructured textual information is elevated from shallow semantic sentiment analysis to deep causal-logical inference. This enables the construction of a logic-aware state space with Bayesian prior characteristics, substantially mitigating the overfitting problem commonly encountered by purely data-driven models under small-sample and extreme-market conditions.

2. From static correlations to evolving topologies in structure modeling: Moving beyond the time-invariant covariance matrix assumption, the proposed logic-driven dynamic graph attention network explicitly captures the nonlinear drift of asset dependency structures in response to evolving market narratives. This provides a topological foundation for identifying and tracing the propagation paths of systemic risk.

3. From expectation maximization to distributionally robust decision-making: Within a stochastic optimal control framework, the integration of IQN-based distributional reinforcement learning and CVaR dual constraints enables full probabilistic modeling of return distributions. Mathematically, this approach is equivalent to searching for an optimal policy frontier under strict worst-case downside risk control, thereby significantly enhancing robustness to fat-tailed return distributions.

5.2. Limitations

Despite its theoretical strengths, the proposed framework faces several practical challenges that warrant careful consideration:

1. Computational complexity and the curse of dimensionality: Dynamic graph construction entails $O(N^2)$ pairwise attention computations. When combined with the inference latency of large language models, this poses substantial computational bottlenecks for full-market, high-frequency trading scenarios involving thousands of assets (N > 4,000).

2. Hallucination risk in large language models: Although DeepSeek-R1 exhibits strong reasoning capabilities, it may still generate spurious or erroneous logic chains when exposed to highly ambiguous or misleading financial rumors. Designing robust hallucination detection and verification mechanisms remains a critical challenge, particularly for future integration with financial knowledge graphs.

3. Local breakdown of stationarity assumptions: While dynamic graphs effectively adapt to gradual distributional drift, they may exhibit delayed structural updates during millisecond-level extreme events such as flash crashes, potentially limiting real-time responsiveness.

5.3. Future research directions

Building upon the above limitations, future research may proceed along the following avenues:

1. Multi-Agent Reinforcement Learning (MARL): Modeling the market as a complex adaptive system composed of heterogeneous agents—retail investors, institutional traders, and market makers—would enable the study of Nash equilibrium strategies of LLM-G-DRL agents under strategic interactions.

2. Offline reinforcement learning: Investigating methods to train robust policies solely from historical static datasets, with strong Out-Of-Distribution (OOD) generalization capabilities, could help circumvent the high costs and risks associated with online trial-and-error learning.

3. Enhanced interpretability: By integrating SHAP values with graph explanation techniques such as GNNExplainer, future work can visualize decision pathways—e.g., "a deterioration in firm A's logic state leads to a reduced allocation to firm B"—thereby strengthening model transparency and trustworthiness in real-world financial applications.

References

[1] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77–91. https: //doi.org/10.2307/2975974

[2] Hong, Y. M., & Wang, S. Y. (2021). Big data, machine learning, and statistics: Challenges and opportunities. *Journal of Econometrics (China), 1*(1), 17–35. https: //doi.org/10.12012/T03-19

[3] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv. https: //doi.org/10.48550/arXiv.1908.10063

[4] DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in large language models via reinforcement learning. arXiv. https: //doi.org/10.48550/arXiv.2501.12948

[5] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. arXiv. https: //doi.org/10.48550/arXiv.2201.11903

[6] Jiang, Z., Xu, D., & Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. arXiv. https: //doi.org/10.48550/arXiv.1706.10059

[7] Xu, B., He, Y. J., Wen, J. C., & Li, X. X. (2024). A review of deep reinforcement learning applications in quantitative trading of financial markets. *Journal of Intelligent Science and Technology, 6*(4), 416–428.

[8] Castillo Pereda, A. I. (2025). Systemic risk and default cascades in global equity markets: Extending the Gai–Kapadia framework with stochastic simulations and network analysis. arXiv. https: //doi.org/10.48550/arXiv.2504.01969

[9] Wang, J., Liao, L., Zhong, K., Deveci, M., du Jardin, P., Tan, J., & Kadry, S. (2024). MRRFGNN: Multi-relation reconstruction and fusion graph neural network for stock crash prediction. *Information Sciences, 678,* 121507. https: //doi.org/10.1016/j.ins.2024.121507

[10] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR). arXiv. https: //doi.org/10.48550/arXiv.1710.10903

[11] Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning (pp. 449–458). PMLR. https: //doi.org/10.48550/arXiv.1707.06887

[12] Dabney, W., Ostrovski, G., Silver, D., & Munos, R. (2018). Implicit quantile networks for distributional reinforcement learning. In Proceedings of the 35th International Conference on Machine Learning (Vol. 80, pp. 1096–1105). PMLR. https: //doi.org/10.48550/arXiv.1806.06923

[13] Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk, 3*, 21–41.