Benchmarking Multimodal Financial Time-Series Forecasting with Self-Evolving Captions

1st Yinjie Teng
School of Computer Science and Technology
East China Normal University
Shanghai, China
51275901082@stu.ecnu.edu.cn

2nd Hongbo Zhao
School of Computer Science and Technology
East China Normal University
Shanghai, China
51275901107@stu.ecnu.edu.cn

***Abstract*—While textual data is increasingly used in financial forecasting, existing multimodal benchmarks rely on static annotations that fail to capture non-stationary market regimes. To address this, we introduce FinEvolve, a scalable framework that reformulates financial captioning as a dynamic, state-adaptive alignment problem. Using multi-agent reasoning and self-evolving optimization, FinEvolve generates faithful, logically consistent, and predictive market captions. Based on this framework, we construct Fin-MM900K, a large-scale dataset of 900K time-series–caption pairs across diverse asset classes, resolutions, and regimes. Furthermore, we establish a comprehensive benchmarking suite for multimodal financial forecasting. Experiments demonstrate that the benefits of textual supervision are highly regime-dependent: it provides substantial gains during crises and structural breaks, but offers limited improvements in noise-dominated markets.**

***Index Terms*—Multimodal Benchmark, Financial Time-Series**

I. Introduction

High-quality textual annotations are pivotal for financial time-series forecasting benchmarks [1]. Unlike general time series, financial markets are driven by heterogeneous information (e.g., news, analyst reports) that encodes early signals invisible to pure numerical models [2]. While recent multimodal approaches integrate text to mitigate price stochasticity [3], [4], effective benchmarking requires these text–time-series pairs to be factually faithful and, crucially, *predictive*. Existing studies often overlook this, treating financial captioning as a static generation task with fixed chart-to-explanation mappings [2]. This conflicts with the *non-stationary* nature of financial markets, where identical price movements under different regimes may stem from distinct causal mechanisms [5].

Furthermore, prevailing datasets rely on the passive scraping of noisy corpora (e.g., StockNet [6], FNSPID [7]) or static LLM prompting without quantitative verification [8]–[10]. Lacking *tool-augmented mechanisms* to enforce cross-modal alignment, these approaches frequently suffer from **causal misalignment** and logical hallucinations. Consequently, they produce generic descriptions that fail to clarify *when and why* text benefits forecasting.

To address these challenges, we propose FinEvolve, transforming unimodal financial time series into *predictive multimodal representations*. Rather than static generation, FinEvolve formulates captioning as a *dynamic, state-adaptive policy optimization problem*. It employs a ReAct-style hierarchical multi-agent system to synthesize technical signals, external news, and historical patterns. A Population-Based Training (PBT) mechanism optimizes agent prompts and tool usage via a hybrid reward, enforcing cross-modal alignment, predictive utility, and logical consistency.

Our main contributions are summarized as follows:

- **Methodology:** We propose FinEvolve, a self-evolving multi-agent framework integrating evolutionary optimization with expert agents for regime-adaptive, predictive financial descriptions.
- **Resource:** We construct Fin-MM900K, comprising 900K high-fidelity time-series–caption pairs. It is the largest instruction-tuned financial dataset explicitly filtered for causal alignment and logical consistency.
- **Benchmarking:** We establish a comprehensive benchmark across 20+ advanced deep learning models, revealing that text-guided forecasting yields significant gains during crises but limited benefits in noise-dominated markets.

II. Methodology: The FinEvolve Framework

A. Problem Formulation and Overview

Financial time-series captioning aims to generate a factually grounded, decision-relevant textual description C for a multivariate market segment $X \in \mathbb{R}^{T\times D}$ (T time steps, D variables). Unlike generic captioning, financial narratives must isolate weak structural signals from stochastic noise and remain robust under non-stationary regimes where the mapping $f : X \to C$ evolves. To address this, we propose FinEvolve (Fig. 1), a self-evolving framework combining a heterogeneous ReAct [11] multi-agent system for structured reasoning with a reinforcement learning (RL)-driven mechanism that continuously adapts agent prompts. High-quality captions require such adaptive, specialized reasoning over static prompting.

B. Heterogeneous Multi-Agent Architecture

FinEvolve formulates caption generation as a collaborative inference process among specialized LLM-based agents:

$$\mathcal{A} = \{A_{\text{quant}}, A_{\text{macro}}, A_{\text{patt}}, A_{\text{narr}}, A_{\text{crit}}\}. \tag{1}$$

TABLE I
AGENT ROLE DEFINITIONS AND COGNITIVE FUNCTIONS. ALL AGENTS ARE BUILT ON GPT-4O.

Agent ($\mathcal{A}_i$)	Role Type	Cognitive Function (f_i)	Key Responsibility
Alpha-Commander	Meta-Controller	$f_{route}(X) \to \{A_{active}\}$	**Regime Identification.** Classifies market states(e.g., *Flash Crash*, *Sideways*) to dynamically activate sub-agents, minimizing redundancy.
Quant-Watcher	Analytical	$f_{tech}(X) \to E_{num}$	**Signal Extraction.** Computes quantitative indicators to mathematically ground the narrative.
News-Hunter	Retrieval (RAG)	$f_{rag}(t, X) \to E_{text}$	**Causal Attribution.** Aligns time t with external knowledge bases to locate exogenous shocks.(e.g.,*Macro events*)
Pattern-Profiler	Memory-Based	$f_{mem}(X, \mathcal{D}_{hist}) \to E_{sim}$	**Historical Analogy.** Applies DTW to retrieve similar past regimes for probabilistic estimates.
Narrator	Synthesis	$f_{gen}(\cup E_i) \to C_{raw}$	**CoT Synthesis.** Aggregates embeddings into a coherent narrative via Chain-of-Thought.
Risk-Critic	Adversarial	$f_{ver}(C_{raw}, X) \to \{0, 1\}$	**Hallucination Check.** Verifies captions against quantitative facts to enforce logical consistency.

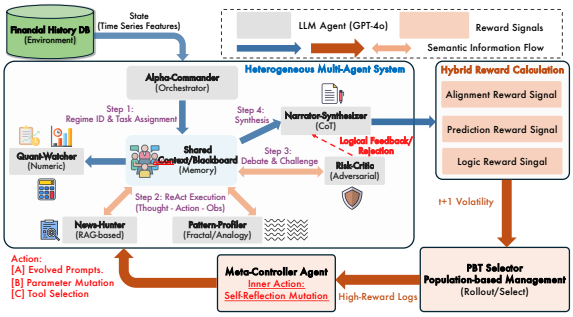

Fig. 1. Overview of our proposed FINEVOLVE framework. An inner ReAct-based multi-agent collaboration solves individual instances, while an outer RL-driven evolutionary loop optimizes agent prompts through a Meta-Controller and hybrid rewards.

Each agent utilizes a dedicated toolset $\mathcal{T}_i$ and system prompt ρ_i (Table I). Following the ReAct paradigm, agents iteratively reason, act, and observe given state X. Specifically, the Quantitative Specialist A_{quant} extracts technical and statistical signals; the Macro Analyst A_{macro} retrieves aligned external context via search; and the Pattern Profiler A_{patt} finds historical analogs using Dynamic Time Warping [12]. Together, their outputs form a shared evidence set $\mathcal{O} = \{o_{\text{quant}}, o_{\text{macro}}, o_{\text{patt}}\}$.

C. Dynamic Collaboration and Risk-Aware Synthesis

Instead of a fixed chain, an Orchestrator A_{orch} dynamically activates agents based on input uncertainty. The Narrator A_{narr} then synthesizes the evidence into a coherent narrative $C_{\text{raw}} = A_{\text{narr}}(X, \mathcal{O})$. To mitigate hallucinations, a Risk Critic A_{crit} evaluates factual alignment, accepting the caption only if it exceeds a consistency threshold τ:

$$C = C_{\text{raw}} \quad \text{s.t.} \quad A_{\text{crit}}(C, X) > \tau. \tag{2}$$

D. Self-Evolving Prompt Optimization

Fixed prompts are brittle under evolving market regimes. We therefore introduce a self-evolving prompt optimization mechanism formulated as a Markov Decision Process (MDP). The state comprises the data batch, prompt population, and market regime. The action space involves mutations by a Meta-Controller LLM, which refines prompts based on performance feedback:

$$\rho_i^{t+1} \leftarrow \text{MetaLLM}(\rho_i^t, \text{Feedback}). \tag{3}$$

To overcome the scarcity of ground-truth captions, we design a hybrid reward optimizing factual correctness and utility:

$$J(\theta) = \mathbb{E}_{\mathcal{P} \sim \pi_\theta} \left[\lambda_1 R_{\text{align}}(C, X) + \lambda_2 R_{\text{pred}}(C) + \lambda_3 R_{\text{logic}}(C) \right], \tag{4}$$

where R_{align} measures semantic consistency via a time-text contrastive embedding, R_{pred} assesses predictive utility via return forecasting, and R_{logic} penalizes inconsistencies detected by the Risk Critic.

Optimization is carried out through Population-Based Training (PBT) [13]. We initialize K agent teams and evaluate them over data mini-batches. High-performing configurations are cloned and mutated by the Meta-Controller based on failure patterns, while underperforming ones are discarded. This evolutionary loop enables FINEVOLVE to continuously adapt its reasoning behaviors to non-stationary market dynamics.

E. The Fin-MM900K Dataset

We propose Fin-MM900K (Fig. 2), a multimodal financial instruction-tuning dataset with 900,000 time-series–caption pairs. These are synthesized via FINEVOLVE to ensure strict grounding of textual descriptions to market signals. Sourced from 14 years (2010–2024) of real-world data (Yahoo Finance, Binance, Bloomberg), the dataset is hierarchically organized across asset classes (equities, crypto, forex, indices, commodities), frequencies (intraday to daily), and market regimes. To address financial market non-stationarity, an Alpha-Commander agent explicitly balances regime samples, over-representing high-impact rare events like market crashes.

Captions generated by the Narrator Agent average 142.5 tokens, featuring a vocabulary of over 18,400 domain-specific terms and 4.2 numerically grounded entities per instance. A Risk-Critic agent strictly filters out 18.4% of initial generations due to hallucinations or inconsistencies, ensuring Fin-MM900K serves as a reliable training corpus and high-integrity benchmark. Table II compares Fin-MM900K with existing datasets [2], [6], [7], [9], [10], [14], [15].

F. Human Expert Evaluation

To assess FINEVOLVE's practical utility, we conducted a double-blind human evaluation with five financial experts (three CFAs and two senior quants). We randomly sampled

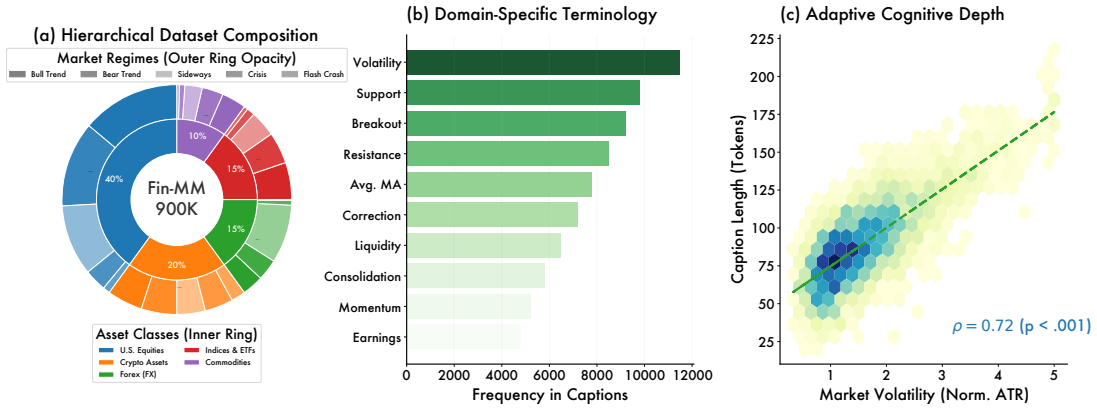

Fig. 2. Overview of the Fin-MM900K dataset: (a) hierarchical composition across asset classes and regimes, (b) frequency of financial terminology, and (c) relationship between volatility and caption length.

TABLE II
COMPARISON WITH EXISTING FINANCIAL MULTIMODAL DATASETS.

Dataset	Venue	Time Span	Granularity	Text Source & Logic	Tools	Alignment	Regime
StockNet [6]	ACL 18	2014-2016	Daily	Raw Tweets (Noisy)	✗	✗	Random
CH-RNN [14]	CIKM 18	2017	Minute	Raw News Headlines	✗	✗	Random
DOW30 [15]	arXiv 23	2020-2022	Daily	Raw Financial News	✗	✗	Bull/Bear
FNSPID [7]	KDD 24	2009-2023	Minute	Raw Global News	✗	✗	Random
TS-FF [10]	EMNLP 24	2010-2020	Quarterly	Earnings Transcripts	✗	✓	Earnings
FinBen [9]	NeurIPS 24	Mixed	Mixed	Aggregated Corpus	✗	✗	Mixed
FinMultiTime [2]	arXiv 25	2009-2025	Minute	Static LLM Caption	✗	✗	Random
Fin-MM900K (Ours)	**-**	**2010-2024**	**Multi-Scale**	**Agent-Evolved (CoT)**	✓	**✓(Critic)**	**Full Coverage**

100 instances across bull, bear, and flash-crash regimes. Annotators evaluated captions from GPT-4o (zero-shot) [39], FinGPT (fine-tuned LLaMA-3 [40]), and FINEVOLVE using a 5-point Likert scale across four dimensions: Faithfulness, Logical Coherence, Professionalism, and Actionable Insight.

As shown in Fig. 3, FINEVOLVE consistently outperforms baselines in Faithfulness and Actionable Insight, indicating more accurate grounding and informative analysis. Although GPT-4o performs competitively in Professionalism, it exhibits lower Faithfulness (e.g., hallucinating unsupported trend reversals)—errors effectively mitigated by FINEVOLVE's critic mechanism. In pairwise comparisons, experts preferred FINEVOLVE over GPT-4o in 68% of cases, primarily citing clearer causal reasoning.

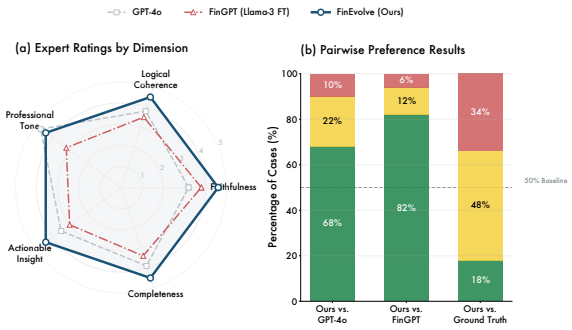

Fig. 3. Human expert evaluation results.

III. BENCHMARKING ON FIN-MM900K

Experimental Setup. *The Text-Guided Adapter Protocol:* To validate dataset universality, we employ a model-agnostic fusion strategy. For any unimodal time series model $\mathcal{M}_{base}$, captions C are encoded via a frozen BERT [41] to yield text embeddings E_{text}. The time series embeddings E_{ts} from $\mathcal{M}_{base}$ act as Queries (Q), while E_{text} serves as Keys (K) and Values (V) in a Multi-Head Cross-Attention layer [25]:

$$E'_{ts} = \text{LayerNorm}(E_{ts} + \text{CrossAttn}(E_{ts}, E_{text}, E_{text})) \quad (5)$$

The enhanced E'_{ts} is then passed to the prediction head of $\mathcal{M}_{base}$, ensuring performance gains stem strictly from text integration. Training hyperparameters (epochs, optimizer, batch size, etc.) strictly follow TSlib [26] defaults. All models are implemented in PyTorch 2.2.2 on 8× NVIDIA A100 (80GB) GPUs.

Baselines & Evaluation Metrics. We evaluate against representative deep forecasting models from TSlib [26], retaining their original configurations unless otherwise specified. Performance is measured via Mean Squared Error (MSE), Mean Absolute Error (MAE), and **Imp.** (defined as the relative MSE reduction achieved by our multimodal approach compared to the unimodal baseline).

A. Main Benchmarking Results

Table III shows that multimodal gains are heterogeneous across architectures. Graph-based models (e.g., TimeFilter) exhibit marginal improvements ($< 0.5\%$), as their rigid structural biases struggle to integrate soft semantic signals. Similarly, simple linear models (e.g., DLinear) benefit little, lacking the capacity for complex numerical-linguistic alignment. Conversely, exogenous-aware architectures (e.g., TimeXer, TFT) demonstrate larger and more consistent gains ($> 1.5\%$) by effectively treating text as structured external covariates. Overall, these results indicate that multimodal fusion is not universally beneficial, it depends critically on architectural flexibility and adaptive context integration.

B. Regime-Specific Analysis

Table IV reveals when multimodal signals are most effective. In sideways markets (dominated by near-Gaussian noise), multimodal integration provides little benefit and can even induce negative transfer in simple linear models (e.g., DLinear, LightTS), as noisy narratives introduce variance rather than

TABLE III

MAIN BENCHMARKING RESULTS ON FIN-MM900K. WE COMPARE UNIMODAL (*Uni*, U) VS. MULTIMODAL (*Multi*, M) PERFORMANCE. VALUES ARE PRESENTED AS U/M. **IMP** DENOTES RELATIVE MSE REDUCTION. **BOLD**: THE BEST RESULTS, UNDERLINE: THE SECOND BEST RESULT. NOTE THAT THE INPUT LENGTH IS FIXED TO 96.

Model	Venue	Horizon 96			Horizon 192			Horizon 336			Horizon 720		
		MSE	MAE	*Imp*	MSE	MAE	*Imp*	MSE	MAE	*Imp*	MSE	MAE	*Imp*
DLinear [16]	AAAI 23	0.481/0.480	0.490/0.489	*0.1%*	0.518/0.516	0.520/0.519	*0.3%*	0.559/0.557	0.552/0.550	*0.4%*	0.601/0.599	0.590/0.589	*0.3%*
TiDE [17]	arXiv 23	0.398/0.396	0.420/0.418	*0.5%*	0.436/0.431	0.454/0.450	*1.1%*	0.476/0.470	0.489/0.484	*1.3%*	0.520/0.514	0.529/0.525	*1.2%*
TimeFilter [18]	ICML 25	**0.334/0.334**	**0.369/0.369**	*0.1%*	**0.370/0.369**	**0.395/0.394**	*0.3%*	**0.407/0.406**	**0.428/0.427**	*0.4%*	**0.457/0.456**	**0.465/0.463**	*0.3%*
Koopa [19]	NeurIPS 23	0.398/0.397	0.421/0.420	*0.3%*	0.438/0.436	0.457/0.456	*0.6%*	0.474/0.471	0.488/0.486	*0.7%*	0.524/0.521	0.531/0.528	*0.7%*
Mamba [20]	COLM 24	0.390/0.387	0.416/0.413	*0.7%*	0.422/0.416	0.442/0.437	*1.4%*	0.464/0.456	0.477/0.471	*1.7%*	0.512/0.504	0.519/0.512	*1.5%*
Autoformer [21]	NeurIPS 21	0.573/0.568	0.563/0.560	*0.7%*	0.602/0.593	0.591/0.585	*1.4%*	0.643/0.632	0.627/0.619	*1.7%*	0.686/0.675	0.666/0.658	*1.6%*
TSMixer [22]	arXiv 23	0.385/0.384	0.410/0.409	*0.2%*	0.429/0.427	0.444/0.442	*0.5%*	0.467/0.464	0.480/0.478	*0.6%*	0.513/0.510	0.523/0.521	*0.6%*
TimeMixer [23]	ICLR 24	0.355/0.353	0.387/0.385	*0.5%*	0.387/0.383	0.415/0.412	*1.1%*	0.425/0.419	0.447/0.442	*1.3%*	0.474/0.468	0.485/0.480	*1.3%*
TFT [24]	arXiv 19	0.584/0.578	0.569/0.565	*0.9%*	0.619/0.608	0.598/0.591	*1.9%*	0.657/0.642	0.630/0.619	*2.3%*	0.704/0.690	0.671/0.661	*2.1%*
Transformer [25]	NeurIPS 17	0.601/0.595	0.581/0.577	*1.0%*	0.636/0.623	0.610/0.601	*2.0%*	0.676/0.660	0.645/0.633	*2.5%*	0.720/0.704	0.686/0.674	*2.2%*
TimesNet [26]	ICLR 23	0.396/0.395	0.419/0.418	*0.3%*	0.434/0.431	0.450/0.448	*0.8%*	0.468/0.465	0.480/0.478	*0.8%*	0.514/0.511	0.519/0.516	*0.7%*
MICN [27]	ICLR 23	0.396/0.394	0.418/0.416	*0.4%*	0.432/0.428	0.453/0.450	*0.8%*	0.474/0.470	0.485/0.482	*0.9%*	0.517/0.512	0.525/0.521	*0.9%*
PatchTST [28]	ICLR 23	0.385/0.382	0.407/0.405	*0.6%*	0.422/0.417	0.441/0.437	*1.2%*	0.461/0.455	0.475/0.470	*1.4%*	0.505/0.498	0.510/0.505	*1.4%*
FEDformer [29]	ICML 22	0.487/0.484	0.495/0.493	*0.5%*	0.527/0.522	0.527/0.523	*1.0%*	0.569/0.562	0.560/0.555	*1.2%*	0.612/0.605	0.599/0.594	*1.1%*
SegRNN [30]	IOTJ 26	0.398/0.397	0.420/0.419	*0.5%*	0.436/0.432	0.455/0.452	*0.9%*	0.476/0.471	0.490/0.486	*1.1%*	0.521/0.516	0.529/0.525	*1.0%*
MultiPatchFormer [31]	SciRep 25	**0.339/0.337**	0.372/0.370	*0.5%*	0.380/0.376	0.407/0.404	*1.0%*	**0.410/0.405**	**0.428/0.424**	*1.2%*	**0.456/0.451**	0.473/0.469	*1.1%*
PAttn [32]	NeurIPS 24	0.358/0.355	0.390/0.387	*0.8%*	0.395/0.389	0.419/0.414	*1.6%*	0.433/0.425	0.455/0.448	*1.9%*	0.480/0.472	0.490/0.483	*1.8%*
TimeXer [33]	NeurIPS 24	0.341/0.337	**0.370/0.367**	*1.1%*	**0.382/0.375**	**0.407/0.401**	*1.9%*	0.417/0.407	0.439/0.431	*2.3%*	0.470/0.461	0.481/0.474	*2.0%*
Crossformer [34]	ICLR 23	0.401/0.398	0.421/0.419	*0.7%*	0.440/0.434	0.455/0.450	*1.4%*	0.476/0.468	0.487/0.481	*1.7%*	0.521/0.513	0.532/0.525	*1.5%*
iTransformer [35]	ICLR 24	0.350/0.348	0.381/0.380	*0.5%*	0.386/0.382	0.413/0.410	*1.1%*	0.425/0.420	0.448/0.443	*1.2%*	0.471/0.466	0.482/0.478	*1.1%*
FreTS [36]	NeurIPS 23	0.407/0.406	0.428/0.428	*0.2%*	0.438/0.437	0.457/0.456	*0.4%*	0.476/0.474	0.488/0.486	*0.5%*	0.522/0.520	0.527/0.525	*0.5%*
WPMixer [37]	AAAI 25	0.345/0.343	0.375/0.373	*0.5%*	0.378/0.375	0.403/0.401	*0.9%*	0.413/0.408	0.435/0.431	*1.1%*	0.461/0.457	**0.471/0.467**	*1.0%*
MSGNet [38]	AAAI 24	0.356/0.355	0.390/0.389	*0.4%*	0.395/0.392	0.421/0.419	*0.7%*	0.429/0.425	0.454/0.451	*1.0%*	0.478/0.474	0.491/0.488	*0.9%*

useful context. In bull trends, numerical models largely capture the dominant momentum, making textual cues auxiliary but consistently helpful.

The most pronounced gains emerge during crisis or collapse regimes. While purely numerical models suffer from inherent lag in reacting to abrupt structural breaks, highly contextual models (e.g., TimeXer, Mamba, Transformer) benefit substantially from textual signals encoding early warnings of extreme events. Importantly, unimodal performance ordering aligns with established benchmarks, confirming that these improvements stem from genuine regime-dependent complementarity rather than artifactually weakened baselines.

Takeaway: When Does Multimodality Help?
Multimodal gains are neither uniform nor guaranteed. Textual signals are most effective for context-aware and exogenous-driven architectures, especially under structural breaks and crisis regimes where numerical models alone suffer from reaction lag. In contrast, rigid graph-based models and low-capacity linear models benefit little and may even exhibit negative transfer in noise-dominated settings. Architectural flexibility and contextual integration ability, rather than multimodality itself, are key prerequisites for successful semantic fusion in financial forecasting.

C. Alignment & Generalization

a) Cross-Modal Retrieval: To evaluate semantic alignment between the financial time-series space $\mathcal{X}$ and textual narrative space $\mathcal{T}$, we conduct bi-directional cross-modal retrieval on the Fin-MM900K test split ($N_{\text{test}} = 10{,}000$). FINEVOLVE representations learned via contrastive pretraining are compared against random retrieval, TS-CLIP [49] (a zero-shot baseline), and FinGPT [50] (LLM hidden states with an MLP encoder). Performance is measured via Recall@K and MRR. Given the similarity of many financial regimes (e.g., sideways markets), R@1 is expected to be moderate, whereas a high R@10 indicates correct matches rank near the top. Notably, text-to-series retrieval is inherently more challenging than series-to-text, as a single narrative may map to multiple plausible price patterns.

b) Zero-Shot Regime Classification: We further investigate whether the learned time-series embeddings support zero-shot market regime classification (e.g., bull, bear, sideways). We compute the cosine similarity between the time-series embedding E_{ts} and textual class prompts (e.g., "This is a bullish trend"), predicting the regime with the highest score. We compare this zero-shot approach against supervised baselines, specifically 10-shot DTW-KNN and a fully supervised ResNet classifier. If the learned semantic alignment is effective, our zero-shot classifier should approach the performance of these supervised methods, demonstrating successful transfer of high-level market semantics into the embedding space.

IV. CONCLUSION

In this paper, we bridge the semantic gap between numerical financial time series and linguistic narratives via FINEVOLVE, a self-evolving multi-agent framework. By formulating financial captioning as dynamic policy optimization rather than a static generation task, we constructed **Fin-MM900K**, the largest and most diverse multimodal financial dataset to date. Extensive benchmarking across 20+ advanced architectures reveals a critical insight: the utility of textual modalities is highly *regime-dependent*. While textual guidance yields marginal gains in noise-dominated sideways markets, it acts as a crucial "safety valve" during structural breaks and crises, capturing exogenous causal drivers invisible to purely numerical models.

TABLE IV
REGIME-SPECIFIC PERFORMANCE ANALYSIS (HORIZON=96). WE EVALUATE ACROSS THREE DISTINCT MARKET REGIMES USING FIN-MM900K.

Model	Venue	**Bull Trend** (Momentum-driven)			**Sideways** (Noise-dominated)			**Crisis / Collapse** (Event-driven)		
		MSE (U/M)	MAE (U/M)	***Imp***	MSE (U/M)	MAE (U/M)	***Imp***	MSE (U/M)	MAE (U/M)	***Imp***
Reformer [42]	ICLR 20	0.605/0.597	0.529/0.525	*1.3%*	0.655/0.650	0.569/0.566	*0.75%*	1.035/0.978	0.662/0.643	*5.5%*
Pyraformer [43]	ICLR 22	0.478/0.475	0.460/0.458	*0.6%*	0.545/0.543	0.505/0.503	*0.35%*	0.958/0.935	0.632/0.624	*2.4%*
PatchTST [28]	ICLR 23	0.345/0.341	0.375/0.373	*1.0%*	0.441/0.440	0.425/0.424	*0.22%*	0.844/0.803	0.591/0.576	*4.9%*
MultiPatchFormer [31]	SciRep 25	0.339/0.337	0.372/0.370	*0.5%*	0.458/0.456	0.435/0.434	*0.35%*	0.805/0.778	0.580/0.569	*3.4%*
ETSformer [44]	arXiv 22	0.458/0.456	0.446/0.444	*0.4%*	0.525/0.523	0.491/0.489	*0.38%*	0.938/0.912	0.625/0.616	*2.8%*
Autoformer [21]	NeurIPS 21	0.555/0.548	0.501/0.497	*1.2%*	0.605/0.602	0.541/0.539	*0.48%*	0.985/0.925	0.642/0.621	*6.1%*
TFT [24]	arXiv 19	0.355/0.351	0.381/0.379	*1.2%*	0.448/0.447	0.428/0.427	*0.16%*	**0.767/0.713**	**0.564/0.543**	*7.1%*
MICN [27]	ICLR 23	0.368/0.366	0.395/0.393	*0.5%*	0.495/0.493	0.468/0.466	*0.35%*	0.895/0.871	0.610/0.601	*2.6%*
TimeXer [33]	NeurIPS 24	0.354/0.349	0.380/0.377	*1.4%*	0.457/0.455	0.432/0.431	*0.31%*	0.908/0.839	0.613/0.589	*7.6%*
TiDE [17]	arXiv 23	0.361/0.359	0.388/0.386	*0.6%*	0.488/0.486	0.462/0.460	*0.41%*	0.885/0.856	0.607/0.596	*3.3%*
WPMixer [37]	AAAI 25	0.345/0.343	0.375/0.373	*0.6%*	0.465/0.463	0.441/0.439	*0.42%*	0.812/0.789	0.584/0.572	*2.8%*
Crossformer [34]	ICLR 23	0.370/0.367	0.397/0.395	*0.8%*	0.498/0.495	0.471/0.469	*0.55%*	0.902/0.865	0.612/0.598	*4.1%*
TimesNet [26]	ICLR 23	0.358/0.357	0.385/0.384	*0.4%*	0.475/0.474	0.451/0.450	*0.15%*	0.865/0.849	0.601/0.595	*1.8%*
LightTS [45]	arXiv 22	0.445/0.444	0.438/0.437	*0.2%*	0.512/0.513	0.480/0.481	*-0.15%*	0.925/0.916	0.620/0.617	*1.0%*
TimeFilter [18]	ICML 25	0.334/0.333	0.369/0.368	*0.3%*	0.453/0.453	0.431/0.431	*0.01%*	0.782/0.771	0.569/0.565	*1.4%*
iTransformer [35]	ICLR 24	0.344/0.341	0.375/0.373	*0.8%*	0.442/0.442	0.425/0.425	*0.08%*	0.825/0.793	0.585/0.573	*3.9%*
DLinear [16]	AAAI 23	**0.329/0.329**	**0.366/0.366**	*0.1%*	**0.429/0.429**	**0.419/0.419**	*-0.02%*	0.898/0.891	0.610/0.607	*0.8%*
Koopa [19]	NeurIPS 23	0.365/0.364	0.392/0.391	*0.2%*	0.491/0.490	0.465/0.464	*0.11%*	0.892/0.881	0.609/0.605	*1.2%*
Mamba [20]	COLM 24	0.334/0.330	0.369/0.367	*1.1%*	**0.431/0.431**	**0.420/0.420**	*0.04%*	**0.776/0.728**	**0.567/0.549**	*6.2%*
Informer [46]	AAAI 21	0.585/0.578	0.518/0.514	*1.2%*	0.635/0.631	0.558/0.555	*0.62%*	1.015/0.958	0.655/0.635	*5.6%*
SCINet [47]	NeurIPS 22	0.472/0.470	0.456/0.454	*0.4%*	0.538/0.537	0.500/0.499	*0.18%*	0.952/0.938	0.630/0.625	*1.5%*
TimeMixer [23]	ICLR 24	0.342/0.340	0.374/0.372	*0.6%*	0.460/0.459	0.437/0.436	*0.28%*	0.815/0.795	0.586/0.576	*2.5%*
FreTS [36]	NeurIPS 23	**0.330/0.329**	**0.367/0.366**	*0.3%*	0.459/0.459	0.433/0.433	*0.05%*	0.781/0.766	0.569/0.563	*1.9%*
FEDformer [29]	ICML 22	0.465/0.461	0.451/0.448	*0.8%*	0.532/0.529	0.496/0.494	*0.45%*	0.945/0.910	0.628/0.615	*3.7%*
Transformer [25]	NeurIPS 17	0.339/0.334	0.371/0.369	*1.5%*	0.450/0.450	0.429/0.429	*0.09%*	0.887/0.812	0.606/0.579	*8.5%*
PAttn [32]	NeurIPS 24	0.355/0.351	0.384/0.381	*1.0%*	0.472/0.470	0.448/0.446	*0.45%*	0.840/0.806	0.595/0.581	*4.1%*
SegRNN [30]	IOTJ 26	0.359/0.357	0.387/0.385	*0.6%*	0.485/0.483	0.459/0.458	*0.32%*	0.875/0.851	0.603/0.594	*2.7%*
TSMixer [22]	arXiv 23	0.362/0.361	0.389/0.388	*0.3%*	0.480/0.481	0.455/0.456	*-0.11%*	0.880/0.868	0.605/0.600	*1.4%*
FiLM [48]	NeurIPS 22	0.485/0.483	0.465/0.463	*0.4%*	0.552/0.551	0.510/0.509	*0.15%*	0.965/0.950	0.635/0.629	*1.6%*

TABLE V
CROSS-MODAL RETRIEVAL PERFORMANCE ON FIN-MM900K.

Model	**Text-to-Series (T2S)**				**Series-to-Text (S2T)**			
	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
Random	0.1	0.5	1.0	0.2	0.1	0.5	1.0	0.2
TS-CLIP (ZS)	6.2	18.5	29.4	11.8	7.5	21.2	33.1	13.5
FinGPT (Emb)	9.4	24.1	38.5	15.2	11.2	28.4	41.6	18.7
FinEvolve	**24.8**	**58.2**	**74.5**	**36.4**	**28.1**	**62.5**	**79.8**	**41.2**

TABLE VI
MARKET REGIME CLASSIFICATION ACCURACY. COMPARISON BETWEEN SUPERVISED METHODS (TRAINED ON LABELS) AND FINEVOLVE (ZERO-SHOT, USING ONLY TEXT EMBEDDINGS).

Model	Type	Bull	Bear	Crisis	Avg. Acc
DTW-KNN	Supervised	72.4	71.5	58.2	67.4
ResNet-1D	Supervised	84.1	83.6	76.5	81.4
TS-CLIP	Zero-shot	55.2	54.8	32.1	47.4
FinEvolve	**Zero-shot**	**81.5**	**80.2**	**79.8**	**80.5**

REFERENCES

[1] P.-D. Arsenault, S. Wang, and J.-M. Patenaude, "A survey of explainable artificial intelligence (xai) in financial time series forecasting," *ACM Computing Surveys*, vol. 57, no. 10, pp. 1–37, 2025.

[2] W. Xu, D. Xiang, Y. Liu, X. Wang, Y. Ma, L. Zhang, S. Hu, C. Xu, and J. Zhang, "Finmultitime: A four-modal bilingual dataset for financial time-series analysis," *arXiv preprint arXiv:2506.05019*, 2025.

[3] H. Liu, S. Xu, Z. Zhao, L. Kong, H. Prabhakar Kamarthi, A. Sasanur, M. Sharma, J. Cui, Q. Wen, C. Zhang *et al.*, "Time-mmd: Multi-domain multimodal dataset for time series analysis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 77 888–77 933, 2024.

[4] Y. Ge, J. Li, Y. Zhao, H. Wen, Z. Li, M. Qiu, H. Li, M. Jin, and S. Pan, "T2s: High-resolution time series generation with text-to-series diffusion models," *arXiv preprint arXiv:2505.02417*, 2025.

[5] H. Yu, J. Dong, and B. Li, "Economic insight through an optimized deep learning model for stock market prediction regarding dow jones industrial average index," *Expert Systems with Applications*, p. 128473, 2025.

[6] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1970–1979.

[7] Z. Dong, X. Fan, and Z. Peng, "Fnspid: A comprehensive financial news dataset in time series," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 4918–4927.

[8] K. J. Koa, Y. Ma, R. Ng, and T.-S. Chua, "Learning to generate explainable stock predictions using self-reflective large language models," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 4304–4315.

[9] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng *et al.*, "Finben: A holistic financial benchmark for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 95 716–95 743, 2024.

[10] R. Koval, N. Andrews, and X. Yan, "Financial forecasting from textual and tabular time series," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 8289–8300.

[11] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The eleventh international conference on learning representations*, 2022.

[12] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd international conference on knowledge discovery and data mining*, 1994, pp. 359–370.

[13] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan *et al.*, "Population based training of neural networks," *arXiv preprint arXiv:1711.09846*, 2017.

[14] H. Wu, W. Zhang, W. Shen, and J. Wang, "Hybrid deep sequential modeling for social text-driven stock prediction," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 1627–1630.

[15] Z. Chen, L. N. Zheng, C. Lu, J. Yuan, and D. Zhu, "Chatgpt informed graph neural network for stock movement prediction," *arXiv preprint arXiv:2306.03763*, 2023.

[16] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 121–11 128.

[17] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu, "Long-term forecasting with tide: Time-series dense encoder," *arXiv preprint arXiv:2304.08424*, 2023.

[18] Y. Hu, G. Zhang, P. Liu, D. Lan, N. Li, D. Cheng, T. Dai, S.-T. Xia, and S. Pan, "Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting," *arXiv preprint arXiv:2501.13041*, 2025.

[19] Y. Liu, C. Li, J. Wang, and M. Long, "Koopa: Learning non-stationary time series dynamics with koopman predictors," *Advances in neural information processing systems*, vol. 36, pp. 12 271–12 290, 2023.

[20] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[21] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22 419–22 430, 2021.

[22] S.-A. Chen, C.-L. Li, N. Yoder, S. O. Arik, and T. Pfister, "Tsmixer: An all-mlp architecture for time series forecasting," *arXiv preprint arXiv:2303.06053*, 2023.

[23] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. Zhou, "Timemixer: Decomposable multiscale mixing for time series forecasting," *arXiv preprint arXiv:2405.14616*, 2024.

[24] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International journal of forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.

[27] H. Wang, J. Peng, F. Huang, J. Wang, J. Chen, and Y. Xiao, "Micn: Multi-scale local and global context modeling for long-term series forecasting," in *The eleventh international conference on learning representations*, 2023.

[28] Y. Nie, "A time series is worth 64words: Long-term forecasting with transformers," *arXiv preprint arXiv:2211.14730*, 2022.

[29] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International conference on machine learning*. PMLR, 2022, pp. 27 268–27 286.

[30] S. Lin, W. Lin, W. Wu, F. Zhao, R. Mo, and H. Zhang, "Segrnn: Segment recurrent neural network for long-term time series forecasting," *IEEE Internet of Things Journal*, 2025.

[31] V. Naghashi, M. Boukadoum, and A. B. Diallo, "A multiscale model for multivariate time series forecasting," *Scientific Reports*, vol. 15, no. 1, p. 1565, 2025.

[32] M. Tan, M. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen, "Are language models actually useful for time series forecasting?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 60 162–60 191, 2024.

[33] Y. Wang, H. Wu, J. Dong, G. Qin, H. Zhang, Y. Liu, Y. Qiu, J. Wang, and M. Long, "Timexer: Empowering transformers for time series forecasting with exogenous variables," *Advances in Neural Information Processing Systems*, vol. 37, pp. 469–498, 2024.

[34] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *The eleventh international conference on learning representations*, 2023.

[35] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.

[36] K. Yi, Q. Zhang, W. Fan, S. Wang, P. Wang, H. He, N. An, D. Lian, L. Cao, and Z. Niu, "Frequency-domain mlps are more effective learners in time series forecasting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 656–76 679, 2023.

[37] M. M. N. Murad, M. Aktukmak, and Y. Yilmaz, "Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 18, 2025, pp. 19 581–19 588.

[38] W. Cai, Y. Liang, X. Liu, J. Feng, and Y. Wu, "Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 10, 2024, pp. 11 141–11 149.

[39] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[40] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[42] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.

[43] S. Liu, H. Yu, C. Liao, J. Li, W. Lin, A. X. Liu, and S. Dustdar, "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=0EXmFzUn5I

[44] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi, "Etsformer: Exponential smoothing transformers for time-series forecasting," *arXiv preprint arXiv:2202.01381*, 2022.

[45] T. Zhang, Y. Zhang, W. Cao, J. Bian, X. Yi, S. Zheng, and J. Li, "Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures," 2022. [Online]. Available: https://arxiv.org/abs/2207.01186

[46] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[47] M. Liu, A. Zeng, M. Chen, Z. Xu, Q. Lai, L. Ma, and Q. Xu, "Scinet: Time series modeling and forecasting with sample convolution and interaction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5816–5828, 2022.

[48] T. Zhou, Z. Ma, Q. Wen, L. Sun, T. Yao, W. Yin, R. Jin *et al.*, "Film: Frequency improved legendre memory model for long-term time series forecasting," *Advances in neural information processing systems*, vol. 35, pp. 12 677–12 690, 2022.

[49] Z. Chen, X. Zhang, and M. Zhu, "Ts-clip: Time series understanding by clip," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 4646–4664.

[50] X.-Y. Liu, G. Wang, H. Yang, and D. Zha, "Fingpt: Democratizing internet-scale data for financial large language models," *arXiv preprint arXiv:2307.10485*, 2023.