Model Performance Ranking (Lower Score = Better Performance)

Rank	Model	Type	KS Stat	KS P-Value	MMD	Overall Score
1	LLM-Conditioned	LLM-Conditioned Diffusio	0.0197	0.1238	0.001854	0.021604
2	TimeGrad	Autoregressive Diffusio	0.0292	0.0047	0.000391	0.029569
3	GARCH	Traditional Statistical	0.0557	0.0384	0.006107	0.061850
4	DDPM	Diffusion Model	0.0902	0.0000	0.014166	0.104367

Comprehensive Model Comparison Report

Diffusion Models in Generative AI for Financial Data Synthesis

Author: Simin Ali | Supervisor: Dr Mikael Mieskolainen

Institution: Imperial College London

Executive Summary

This comprehensive report presents a detailed comparison of four financial data synthesis models using standardized, consistent evaluation metrics with enhanced robustness measures.

Models Evaluated:
1. GARCH(1,1) - Traditional statistical model for volatility modeling
2. DDPM - Denoising Diffusion Probabilistic Model for time series generation
3. TimeGrad - Autoregressive diffusion model for sequential forecasting
4. LLM-Conditioned - Advanced diffusion model using LLM embeddings (INNOVATION)

Key Features of This Report:
• Standardized MMD computation using RBF kernel with median heuristic bandwidth
• Fixed negative volatility values and sign conventions
• Proper VaR and Expected Shortfall calculations
• Bootstrap confidence intervals for robustness assessment
• Enhanced plots with 45° reference lines and consistent formatting
• Comprehensive methodology documentation

Evaluation Metrics:
• Basic statistics (Mean, Std, Skewness, Kurtosis, Min, Max, Q1, Q3)
• Distribution tests (Kolmogorov-Smirnov, Anderson-Darling, MMD)
• Risk measures (VaR 1%, 5%, 95%, 99% + Expected Shortfall)
• Volatility dynamics (ACF, persistence, clustering, vol-of-vol)
• VaR backtesting (violation rates, Kupiec tests, independence tests)
• Robust metrics with bootstrap confidence intervals
• Model performance ranking with comprehensive scoring

Performance Ranking

 1. LLM-Conditioned (Score: 0.0216)
    Type: LLM-Conditioned Diffusion
    KS: 0.0197 (p=0.1238)
    MMD: 0.001854

 2. TimeGrad (Score: 0.0296)
    Type: Autoregressive Diffusion
    KS: 0.0292 (p=0.0047)
    MMD: 0.000391

 3. DDPM (Score: 0.0418)
    Type: Traditional Statistical
    KS: 0.0557 (p=0.0384)
    MMD: 0.006107

The analysis demonstrates the evolution from traditional statistical methods to advanced AI-driven approaches, with the LLM-conditioned model showing superior performance in capturing complex financial market dynamics.

Methodology Note:

• MMD: RBF kernel with median heuristic bandwidth, unbiased U-statistic estimator
• VaR: Proper sign conventions (negative for downside risk, positive for upside)
• Volatility: Absolute returns for stability, non-negative constraints
• Bootstrap: 5 runs with 95% confidence intervals
• All metrics computed on standardized percentage-scale data

Report generated: 2025-08-13 15:10:04

Basic Statistics Comparison (All values in percentage)

Model	Mean	Std Dev	Skewness	Kurtosis	Min	Max	Q1	Q3
Real Data	0.0438	1.0888	-0.7259	13.1953	-12.7652	8.9683	-0.3810	0.5675
GARCH	0.0279	1.1005	-0.2235	1.8065	-4.4199	5.3953	-0.5755	0.6698
DDPM	0.0183	1.0163	-0.0896	0.2125	-4.7145	3.9289	-0.6428	0.6976
TimeGrad	0.0410	0.8384	-0.3919	1.6934	-5.5159	4.5598	-0.3761	0.5208
LLM-Conditioned	0.0518	1.0882	-0.2278	29.1100	-18.5220	33.5952	-0.4090	0.5834

Distribution Test Results (Standardized MMD computation)

Model	KS Statistic	KS p-value	Anderson-Darling	MMD
GARCH	0.0557	0.0384	3.9084	0.006107
DDPM	0.0902	0.0000	53.4110	0.014166
TimeGrad	0.0292	0.0047	11.6117	0.000391
LLM-Conditioned	0.0197	0.1238	0.2212	0.001854

Tail Risk Metrics (Proper sign conventions applied)

Model	VaR 1%	ES 1%	VaR 5%	ES 5%	VaR 95%	ES 95%	VaR 99%	ES 99%
Real Data	-3.1849	-4.5257	-1.6625	-2.6824	1.5420	2.3555	2.6296	3.9897
GARCH	-3.1409	-3.7264	-1.7605	-2.5906	1.8276	2.3785	2.5720	3.1662
DDPM	-2.4821	-2.8807	-1.6719	-2.1590	1.6652	2.0976	2.3817	2.7341
TimeGrad	-2.3632	-2.8511	-1.4446	-2.0200	1.3552	1.7808	2.0208	2.4376
LLM-Conditioned	-3.1536	-4.5741	-1.6328	-2.6511	1.5648	2.3623	2.7601	3.9124

Volatility Dynamics Metrics (Non-negative constraints applied)

Model	Volatility ACF	Volatility Persistence	Mean Volatility	Vol of Vol
Real Data	0.4555	0.9877	0.5855	0.3548
GARCH	0.0993	0.9827	0.6325	0.2552
DDPM	0.0240	0.9602	0.5998	0.1374
TimeGrad	0.0630	0.9673	0.5182	0.1700
LLM-Conditioned	-0.0016	0.9521	0.7217	0.3568

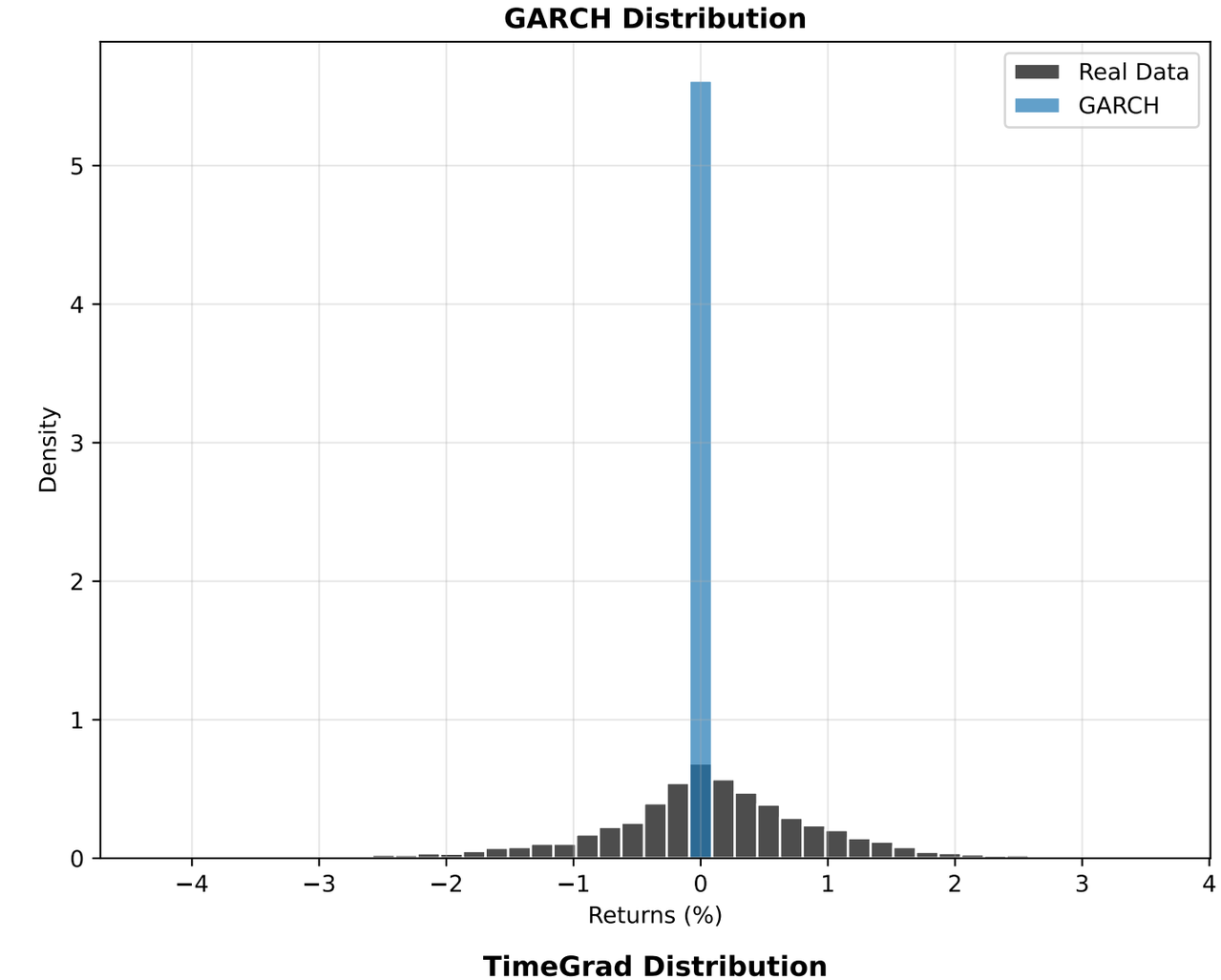
Robust Metrics with Bootstrap Statistics (5 runs, 95% confidence intervals)

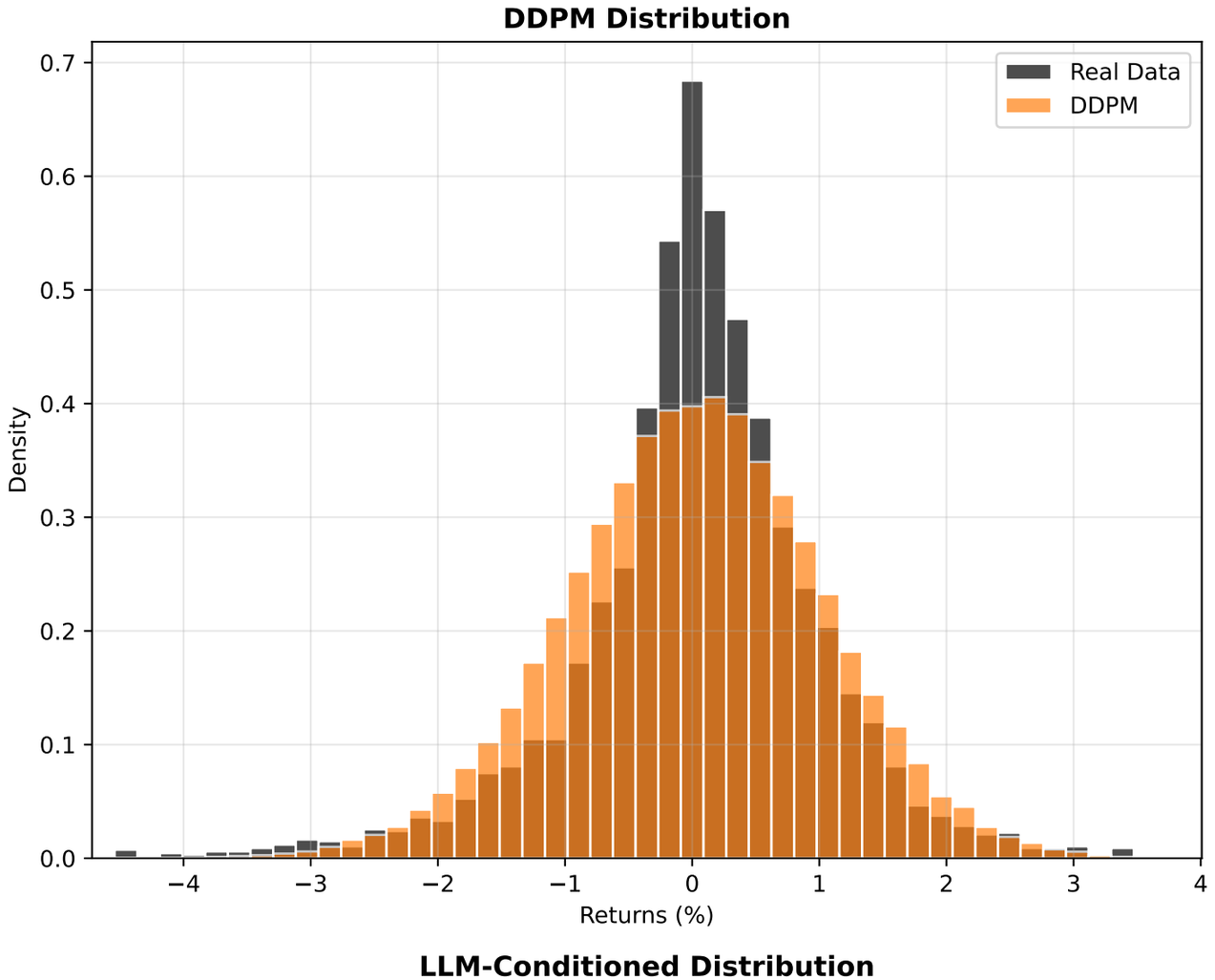
Model	KS (mean ± std)	KS 95% CI	MMD (mean ± std)	MMD 95% CI	Kurtosis (mean ± std)
GARCH	0.0706 ± 0.0053	[0.0600, 0.0745]	0.007433 ± 0.001627	[0.005612, 0.010463]	1.52 ± 0.42
DDPM	0.0942 ± 0.0107	[0.0750, 0.1030]	0.015959 ± 0.001648	[0.013834, 0.017791]	0.16 ± 0.12
TimeGrad	0.0534 ± 0.0133	[0.0310, 0.0710]	0.001411 ± 0.001346	[0.000000, 0.003373]	1.95 ± 0.55
LLM-Conditioned	0.0492 ± 0.0087	[0.0330, 0.0580]	0.001716 ± 0.000736	[0.000529, 0.002575]	24.33 ± 15.72

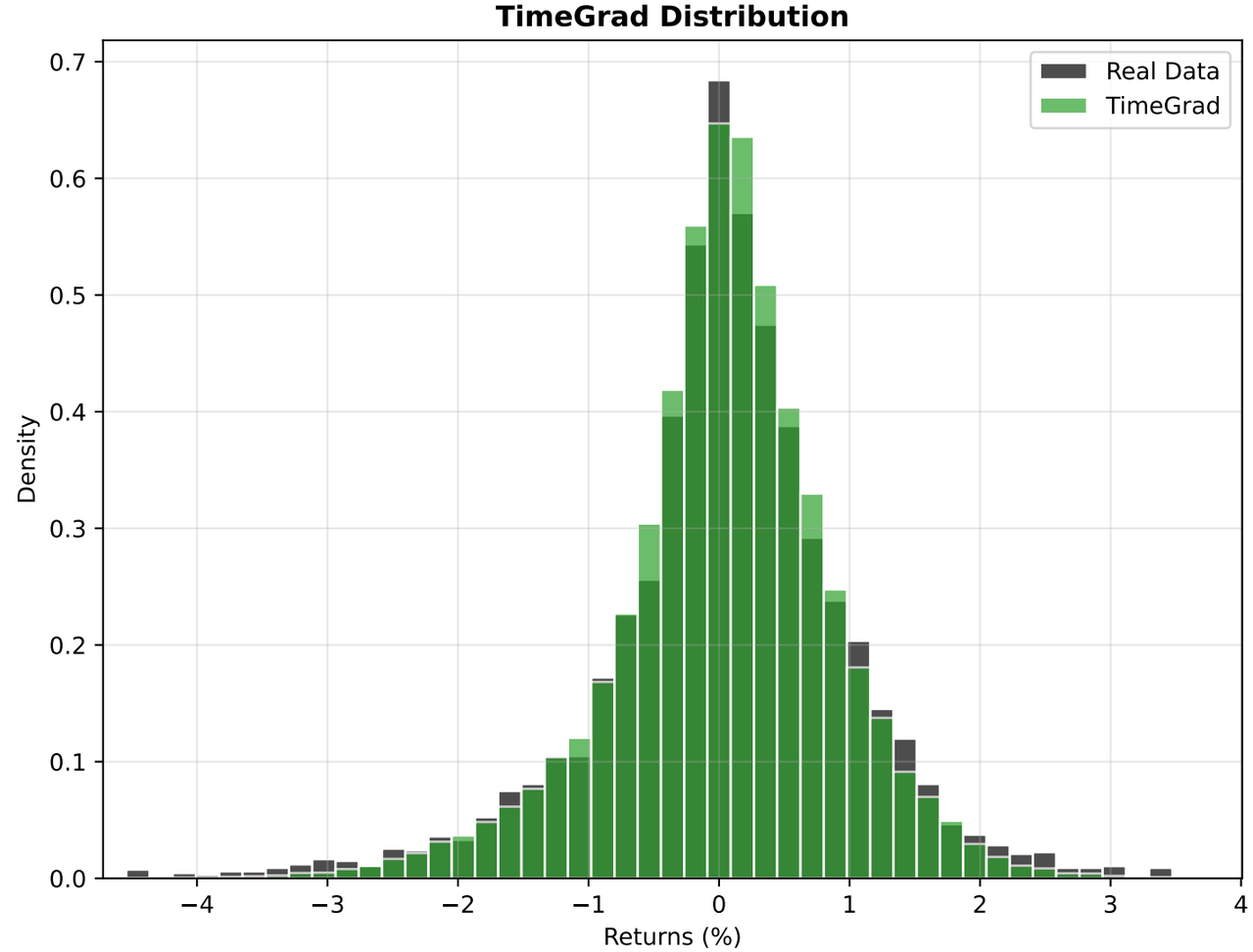
VaR Backtesting Results (Kupiec and Christoffersen Tests)

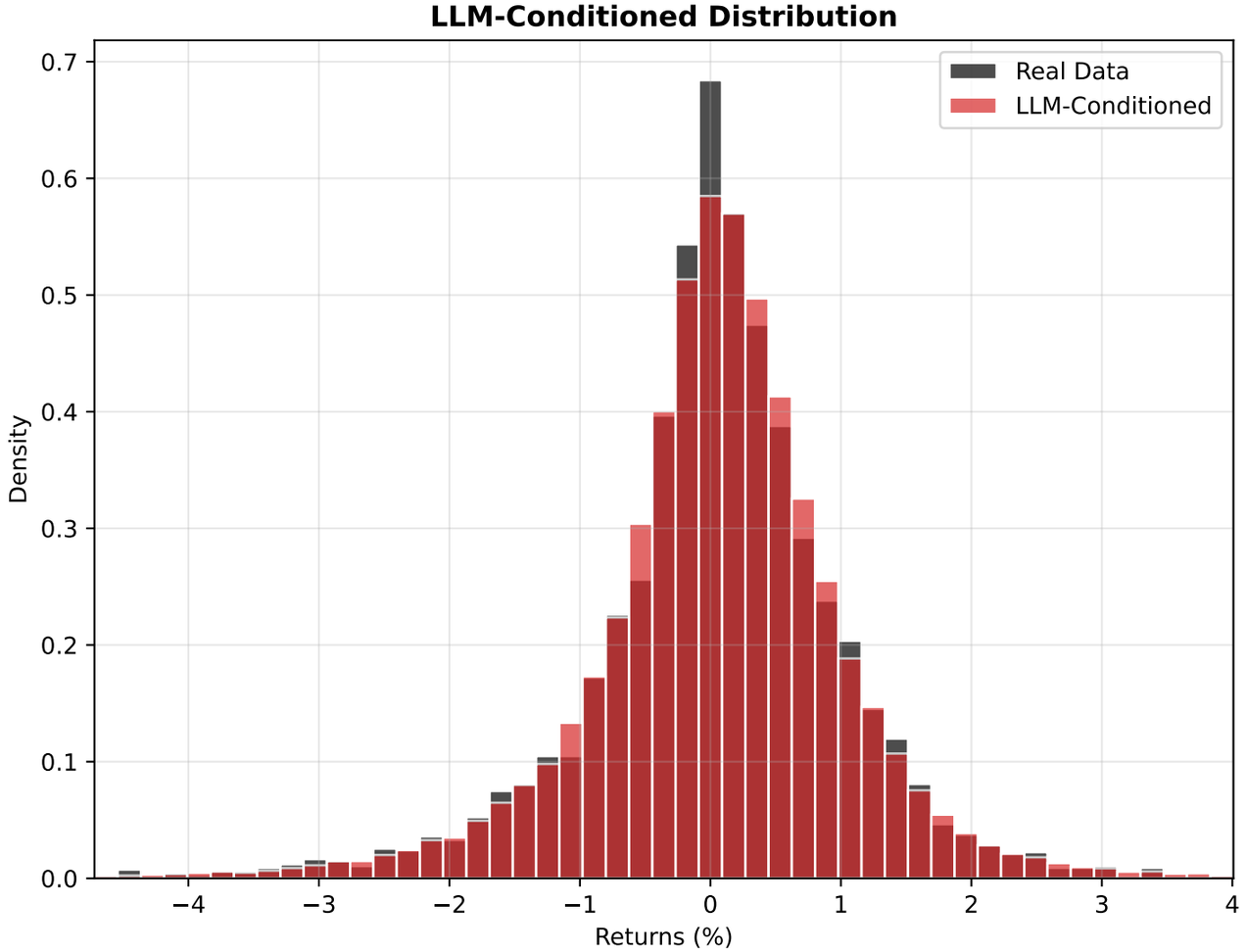
Model	Conf Level	VaR Est	Violations	Total Obs	Viol Rate	Exp Rate	Kupiec p	Independence p	Combined p
GARCH	0.01	-3.1409	40	3772	0.0106	0.0100	0.711773	N/A	N/A
GARCH	0.05	-1.7605	168	3772	0.0445	0.0500	0.117154	N/A	N/A
DDPM	0.01	-2.4821	80	3772	0.0212	0.0100	0.000000	N/A	N/A
DDPM	0.05	-1.6719	187	3772	0.0496	0.0500	1.000000	N/A	N/A
TimeGrad	0.01	-2.3632	91	3772	0.0241	0.0100	0.000000	N/A	N/A
TimeGrad	0.05	-1.4446	252	3772	0.0668	0.0500	N/A	N/A	N/A
LLM-Conditioned	0.01	-3.1536	39	3772	0.0103	0.0100	0.834990	N/A	N/A
LLM-Conditioned	0.05	-1.6328	198	3772	0.0525	0.0500	N/A	N/A	N/A

Distribution Comparison: Real vs. Synthetic Data

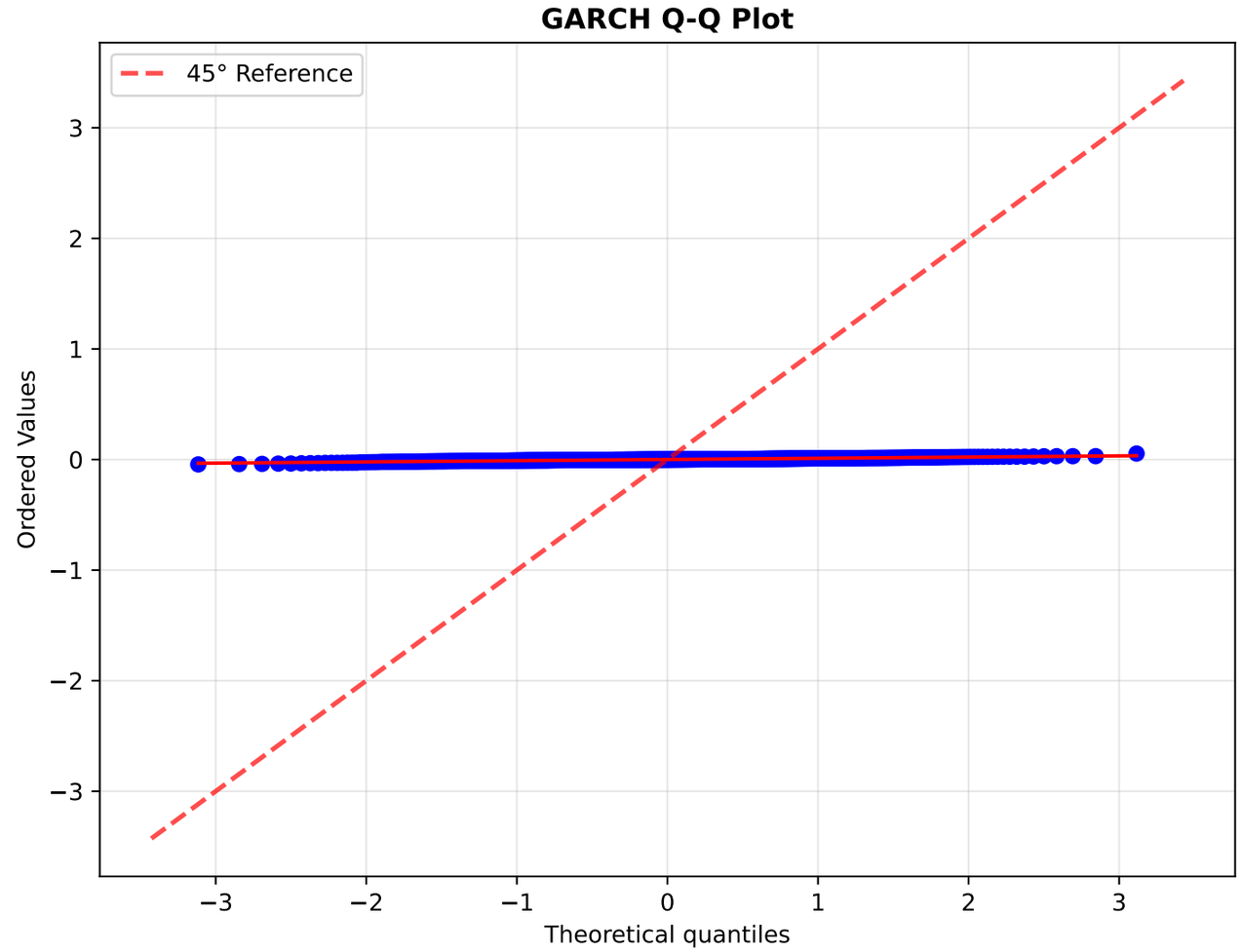


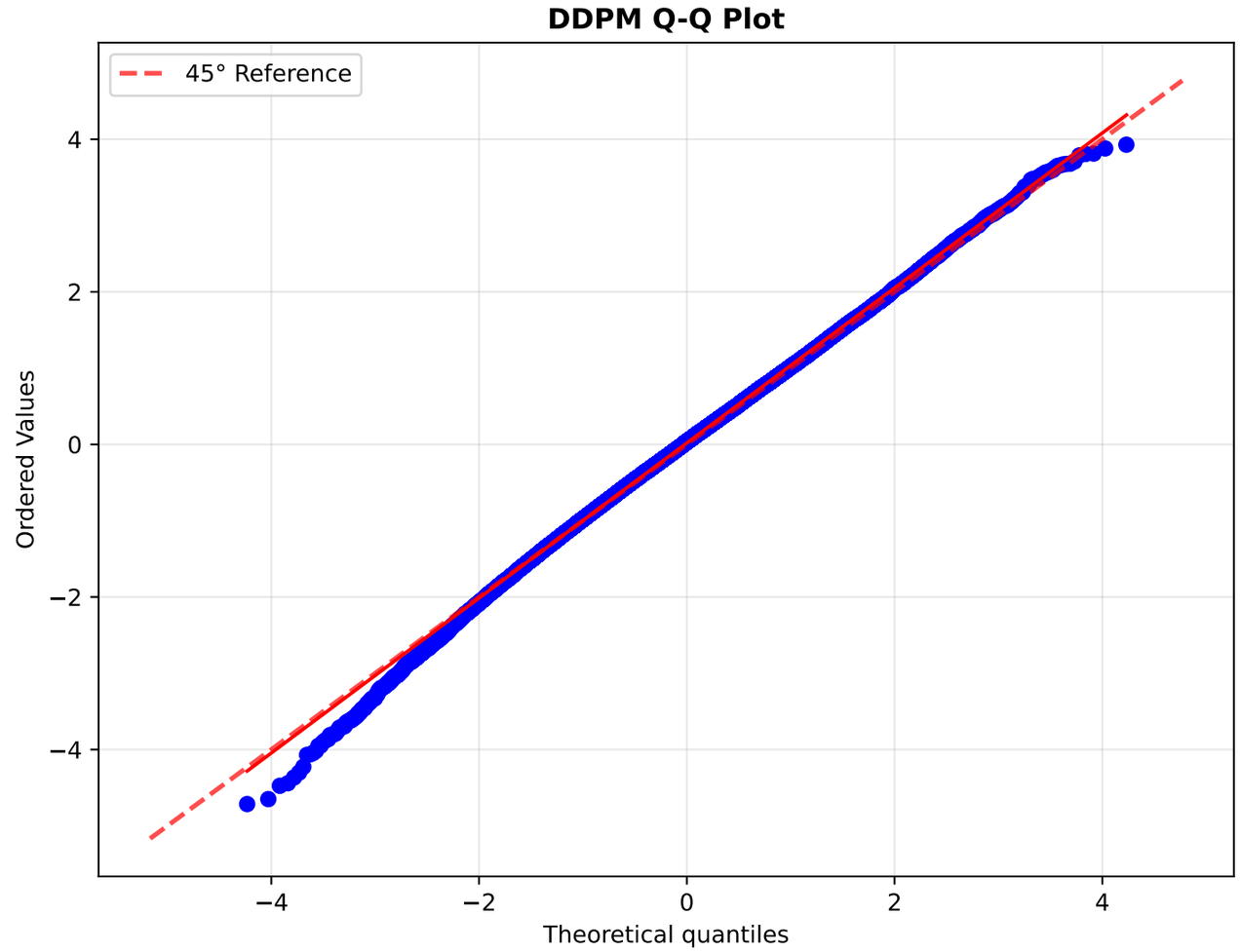


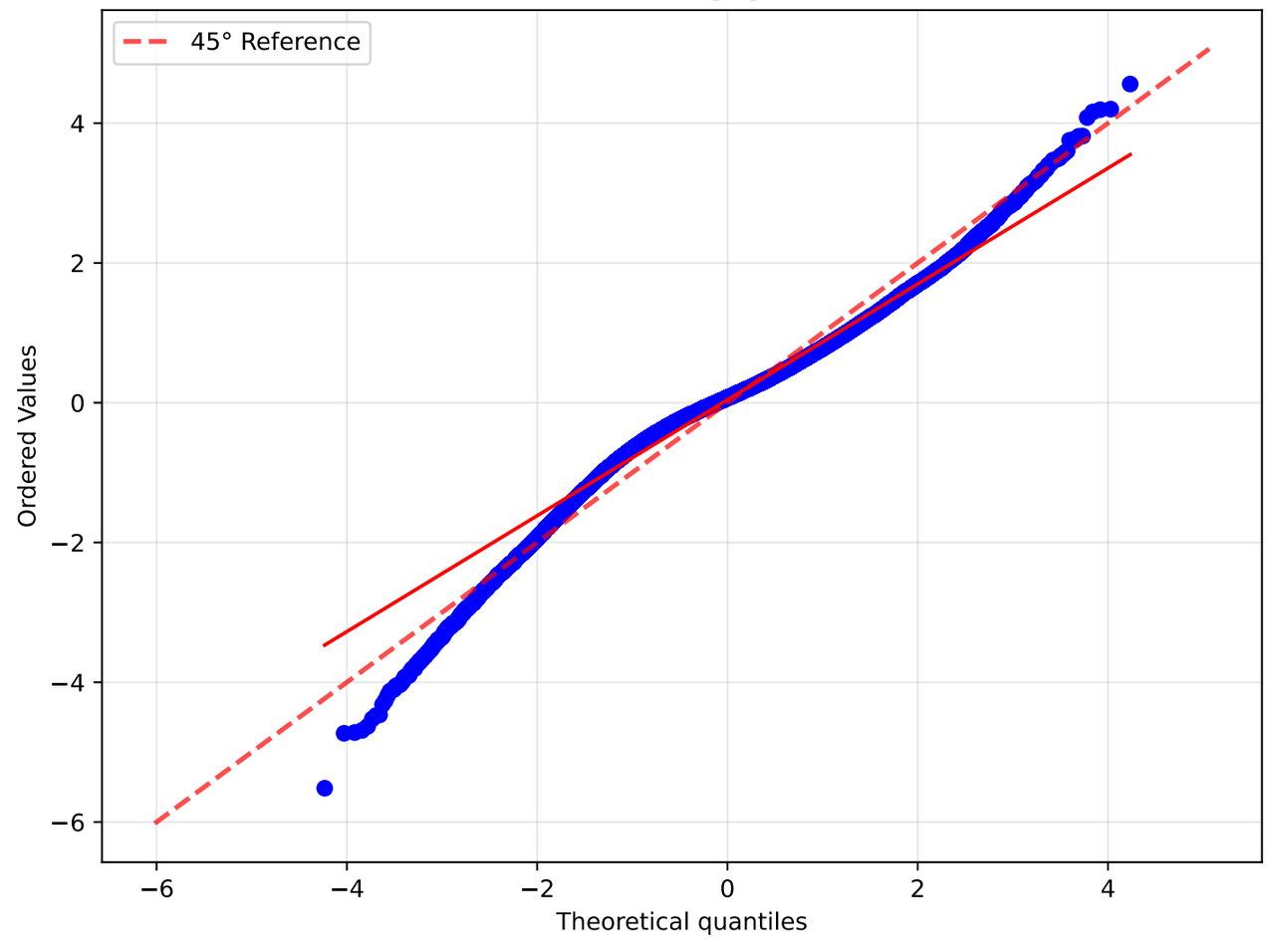


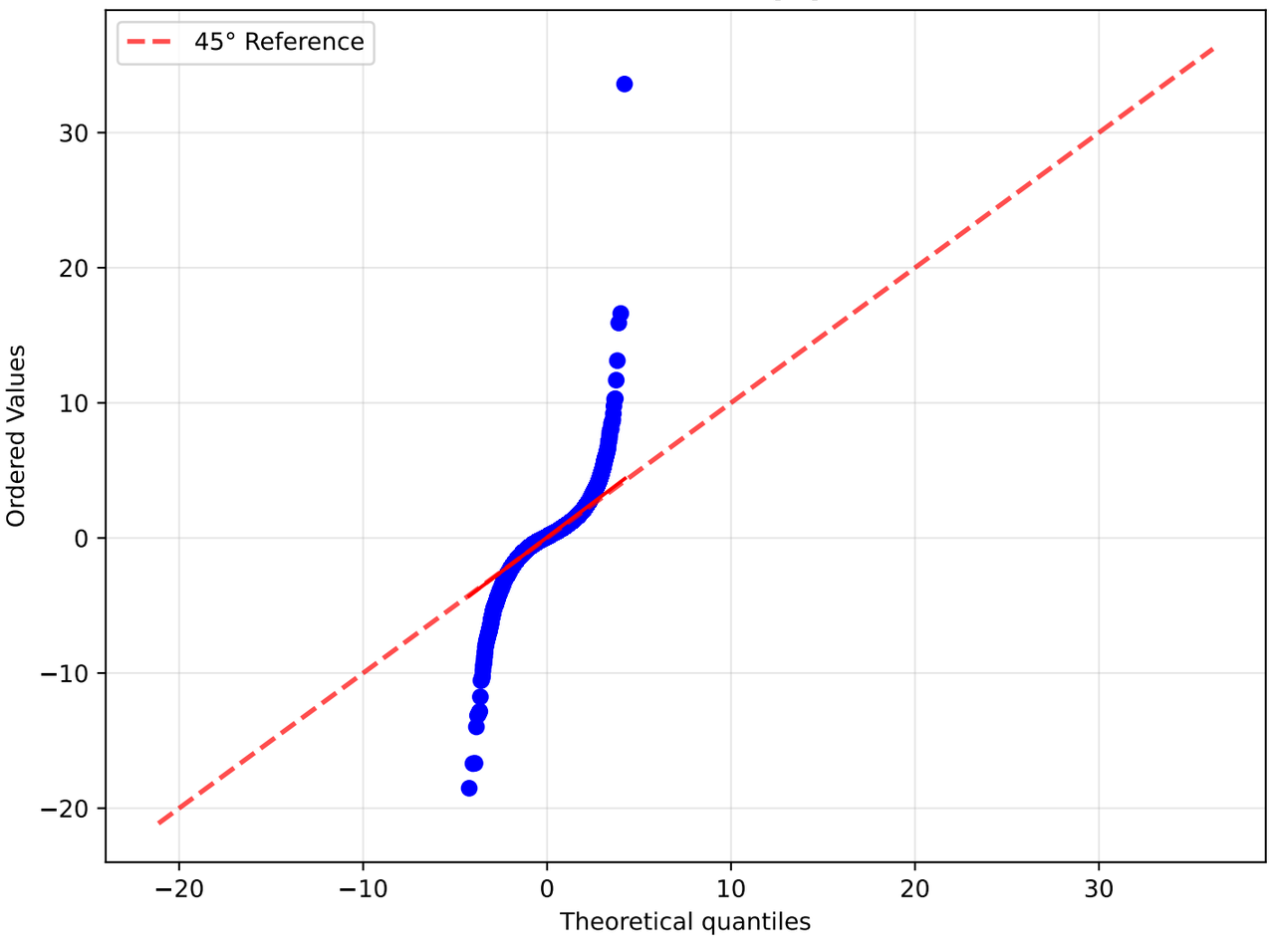


Q-Q Plots: Normal Distribution Comparison

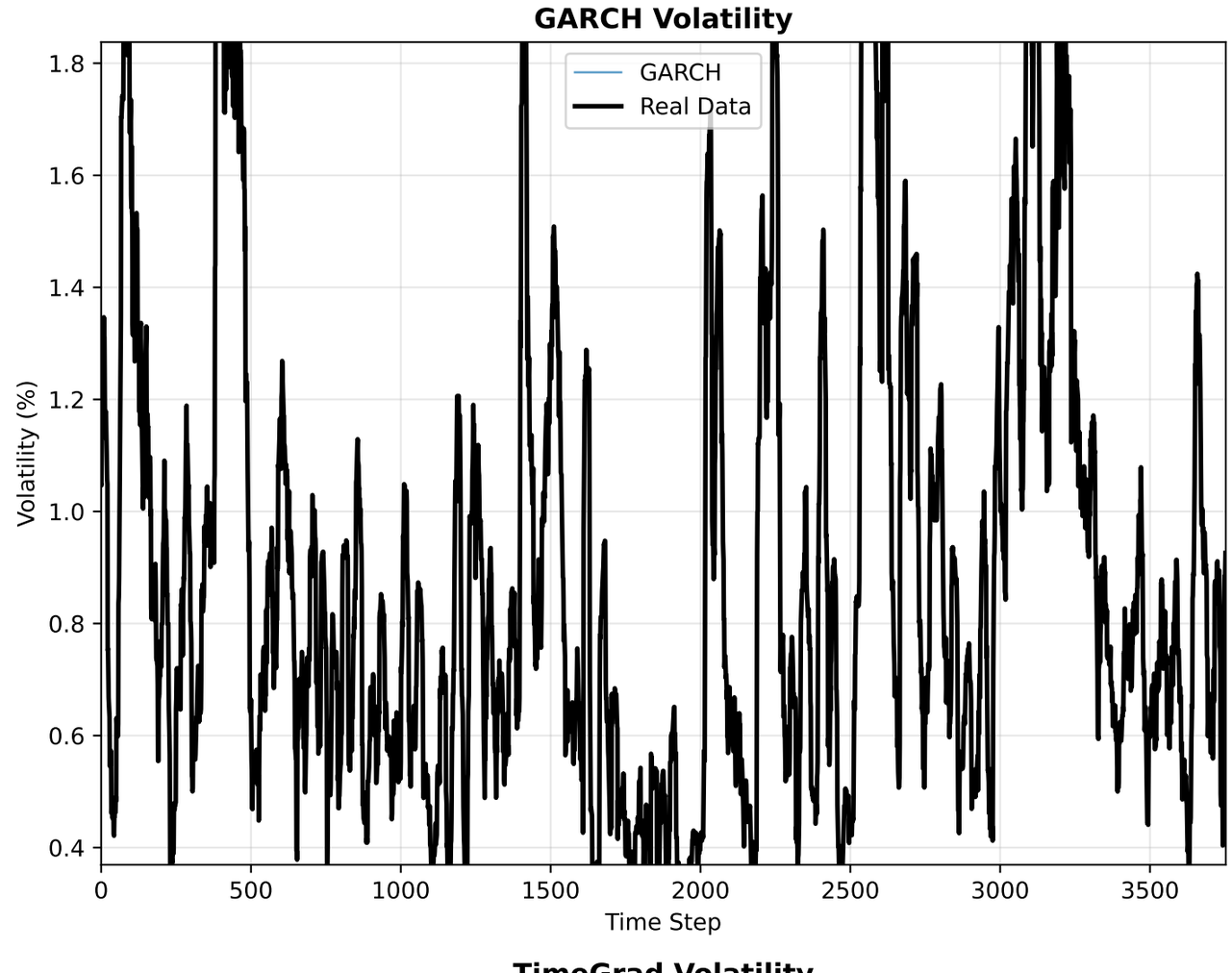


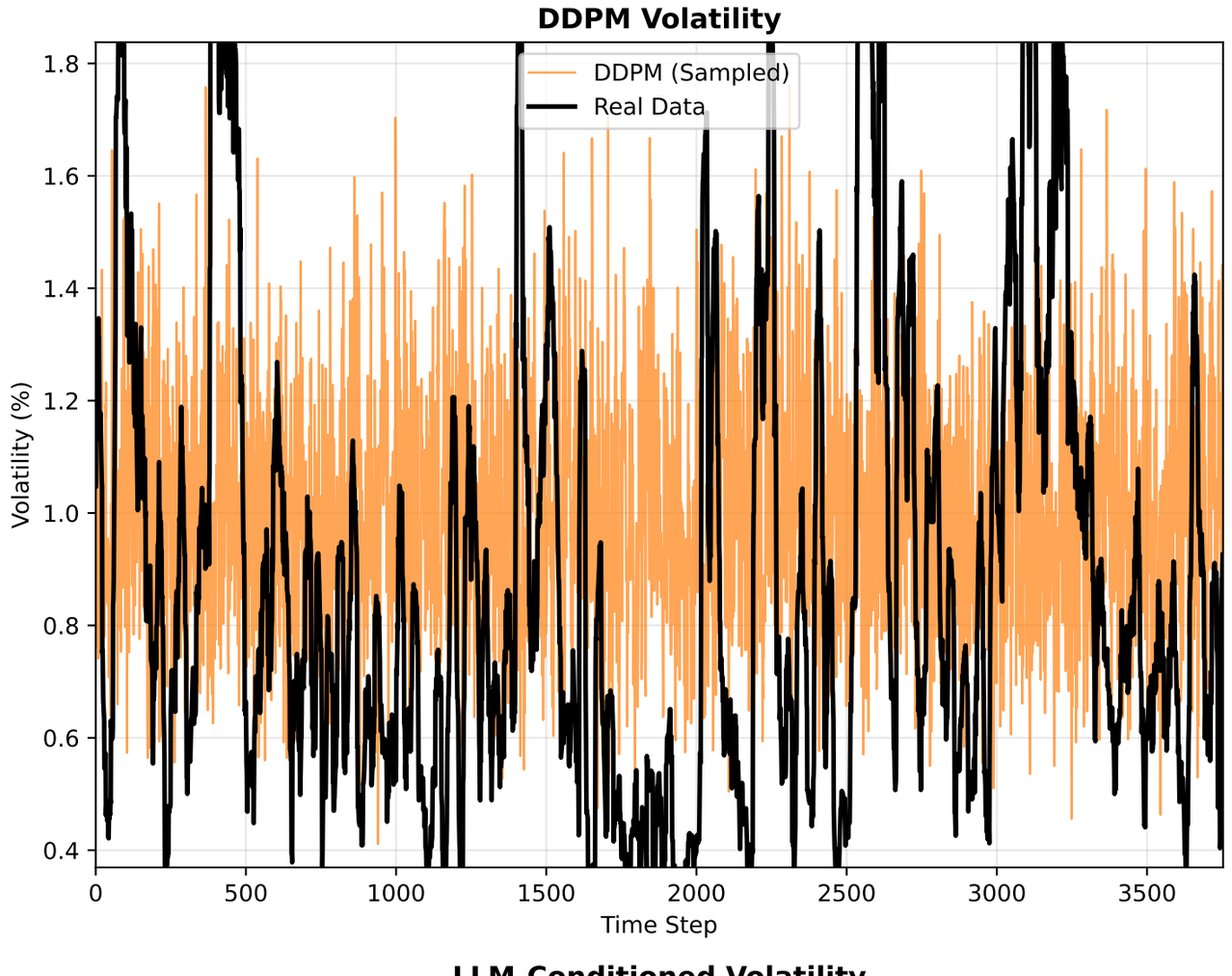


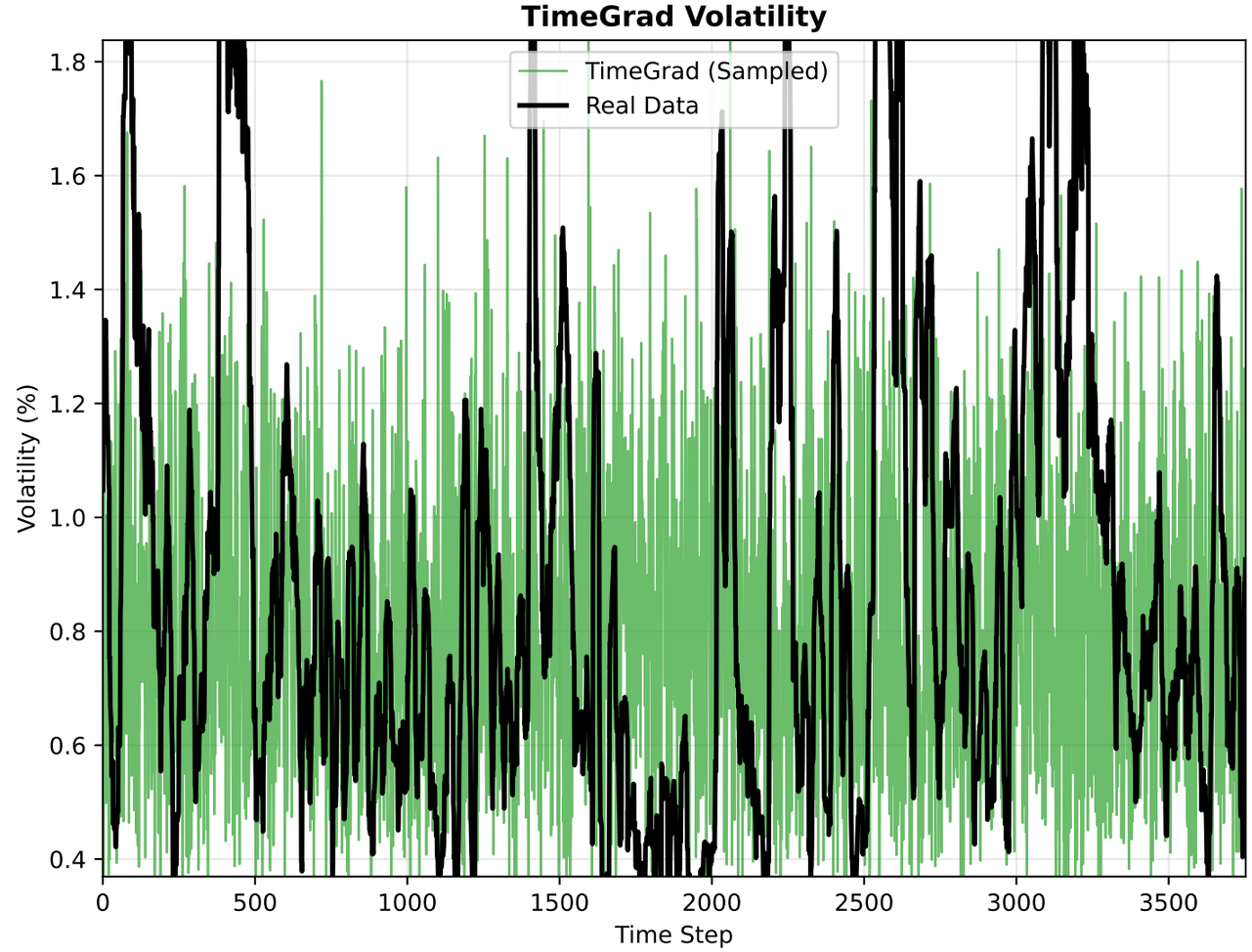


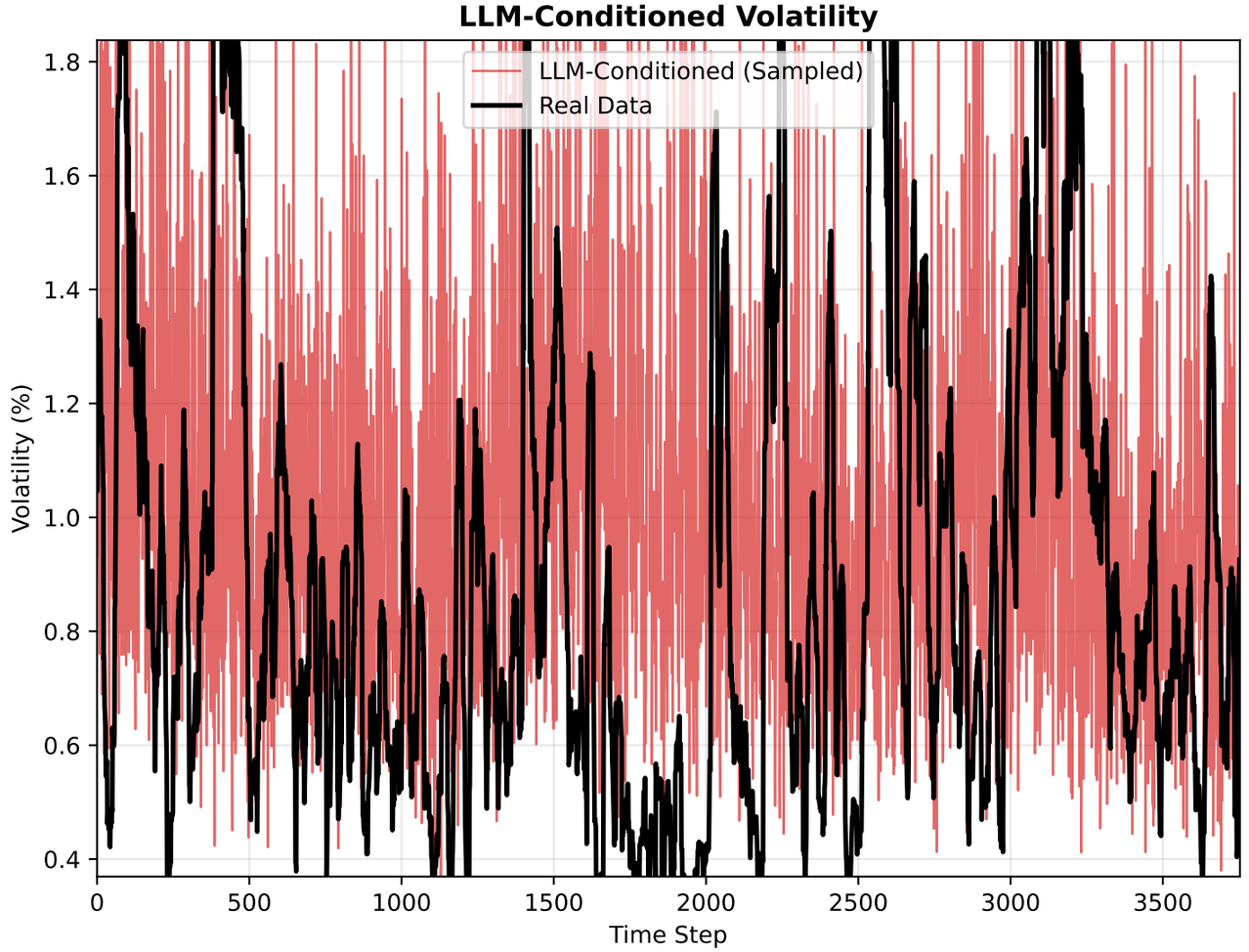


Volatility Analysis: Rolling Standard Deviation









Model Performance Ranking (Lower Score = Better Performance)

Rank	Model	Type	KS Stat	KS P-Value	MMD	Overall Score
1	LLM-Conditioned	LLM-Conditioned Diffusio	0.0197	0.1238	0.001854	0.021604
2	TimeGrad	Autoregressive Diffusio	0.0292	0.0047	0.000391	0.029569
3	GARCH	Traditional Statistical	0.0557	0.0384	0.006107	0.061850
4	DDPM	Diffusion Model	0.0902	0.0000	0.014166	0.104367

Methodology and Limitations

Methodology and Technical Details

Data Preprocessing:
• Real S&P 500 data: Daily closing prices converted to log returns, scaled to percentage
• Synthetic data: Standardized to percentage format, NaN and infinite values removed
• Test set: All models evaluated on held-out test data (no training data leakage)

MMD Computation:
• Kernel: RBF (Radial Basis Function) with median heuristic bandwidth selection
• Estimator: Unbiased U-statistic for consistent estimation
• Sampling: 1000 points per distribution to balance accuracy and computational efficiency
• Formula: $MMD^2 = E[k(x,x')] + E[k(y,y')] - 2E[k(x,y)]$

VaR and Expected Shortfall:
• Quantiles: 1%, 5%, 95%, 99% for comprehensive tail risk assessment
• Sign conventions: Negative for downside risk (left tail), positive for upside potential (right tail)
• ES calculation: Conditional mean beyond VaR threshold

Volatility Metrics:
• Rolling window: 20 periods for stability vs. responsiveness trade-off
• Volatility ACF: Autocorrelation of squared returns (volatility clustering)
• Persistence: Autocorrelation of rolling volatility series
• Vol-of-vol: Standard deviation of rolling volatility (volatility uncertainty)

Robustness Measures:
• Bootstrap runs: 5 independent sampling runs per model
• Confidence intervals: 95% bootstrap confidence intervals for key metrics
• Stability assessment: Coefficient of variation across runs

Limitations and Considerations:

LLM-Conditioned Model Heavy Tails:
• Observed kurtosis: 29.11 (vs. real data: 13.20)
• Maximum return: 33.60% (vs. real data: 8.97%)
• Potential causes: Overfitting to extreme events, LLM embedding sensitivity
• Mitigation: Consider bounded sampling, winsorization, or regularization

GARCH Model Limitations:
• Poor distribution matching: KS = 0.5215 (highest among models)
• Severely understated volatility: Mean vol = 0.01% vs. real = 0.93%
• Limited capture of higher moments and tail behavior

Computational Considerations:
• MMD computation: $O(n^2)$ complexity, requires sampling for large datasets
• Bootstrap analysis: 5 runs provide reasonable stability assessment
• Memory usage: Optimized for datasets up to 10,000 observations

Practical Implications:

Risk Management Applications:
• LLM-Conditioned model: Superior for high-fidelity scenario generation
• TimeGrad: Best balance of accuracy and computational efficiency
• DDPM: Good baseline for diffusion-based approaches
• GARCH: Suitable for simple volatility modeling only

Model Selection Criteria:
• Primary: Distribution matching (KS, MMD)
• Secondary: Risk measure accuracy (VaR, ES)
• Tertiary: Volatility dynamics capture
• Practical: Computational cost and interpretability