Diffusion-Based Probabilistic Forecasting for Financial Time Series: A Walk-Forward Study

Lorenzo Price, Gavin Leema and Andrew Collado
Group #: 7
Department of Computer Science and Engineering
University at Buffalo
Buffalo, NY 14203
{lorenzop;gmleema;arcollad}@buffalo.edu

Abstract

Financial time series forecasting remains challenging due to inherent volatility (price variance), non-stationarity (changing statistical properties over time), and extreme tail events (market crashes). We present a diffusion-based generative model for probabilistic forecasting of stock market returns, specifically targeting the S&P 500 ETF (SPY). Our approach employs Denoising Diffusion Probabilistic Models (DDPMs) to generate multiple future return trajectories. This allows us to capture both aleatoric uncertainty (intrinsic market noise) and epistemic uncertainty (model uncertainty). We evaluate our model using rigorous walk-forward cross-validation—a method that simulates real-world trading by never showing the model future data—across 14 years including the 2008 financial crisis. Results show competitive performance with a mean CRPS of 0.0465. While our model performs comparably to classical baselines like GARCH in calm periods, we provide critical analysis of its behavior during market crises.

1 Introduction

Financial markets exhibit complex dynamics characterized by *volatility clustering* (periods of calm are followed by periods of high variance) and *heavy-tailed distributions* (extreme events occur more often than a normal "Bell Curve" predicts) [3]. Accurate probabilistic forecasts are crucial for risk management—specifically for estimating metrics like Value-at-Risk (VaR), which quantifies the potential loss in a portfolio.

Traditional approaches make strong assumptions. For instance, the GARCH model [4] assumes volatility follows a specific mathematical decay, which may not capture the full complexity of modern algorithmic markets. Recent advances in deep generative models, particularly diffusion models, offer a promising alternative by learning these complex distributions directly from data without restrictive assumptions.

1.1 Our Contributions

We provide a comprehensive evaluation of diffusion models for financial time series:

1. **Novel Application**: Evaluation of DDPM-based forecasting for equity returns using rigorous walk-forward cross-validation spanning 25 years and major market regimes.
2. **Methodological Rigor**: Implementation of proper time series validation preventing *data leakage* (the accidental use of future information), a common pitfall in financial ML.
3. **Comprehensive Evaluation**: Multi-faceted assessment using proper scoring rules (CRPS), calibration analysis (PIT histograms), and distributional comparisons.

4. **Practical Insights**: Demonstration that diffusion models achieve competitive performance (CRPS: 0.0465) with well-calibrated uncertainty (81.9% coverage) while revealing limitations during crisis periods.

2 Related Works

Generative Models in Finance. Diffusion models, originally designed for image generation, have gained attention for time series. They work by destroying data with noise and learning to reverse the process to generate new samples. Meijer and Chen [5] survey diffusion for forecasting, while Briazkalo [1] apply them to financial series. Our work extends these approaches by implementing rigorous walk-forward validation instead of random train-test splits, which are invalid for time-series data.

Classical Econometric Baselines. To benchmark our Deep Learning model, we use two industry standards:

- **GARCH (Generalized Autoregressive Conditional Heteroskedasticity)** [4]: The gold standard for volatility forecasting. It models the "clustering" of volatility where large price changes tend to follow large price changes.
- **AR (Autoregressive) Models** [2]: Simple linear models that predict future returns based on a weighted sum of past returns.

We demonstrate that diffusion offers comparable accuracy to these established methods while providing more flexible uncertainty quantification.

3 Data

We utilize daily closing prices of the SPDR S&P 500 ETF Trust (ticker: SPY) spanning January 4, 2000 to November 26, 2025 (6,515 trading days). SPY tracks the S&P 500 index and represents a diversified portfolio of large-cap U.S. equities. Data was obtained via the `yfinance` Python library.

3.1 Preprocessing

We compute **log returns** rather than raw prices. Log returns are preferred in finance because they are time-additive and often more stationary (statistically stable) than raw prices:

$$r_t = \log(P_t/P_{t-1}) \tag{1}$$

Returns are then **z-score normalized** (scaled to have mean 0 and variance 1) within each training fold. **Critical Note**: The scaler is fitted *only* on training data and applied to validation/test sets. This prevents information leakage, ensuring the model cannot "see" the range of future prices.

Table 1 summarizes the data statistics. The **excess kurtosis** of 10.8 is significant; a normal distribution has a kurtosis of 0. This confirms the presence of "fat tails"—extreme market moves are far more common than standard statistics would suggest.

Table 1: SPY returns summary statistics (2000-2025)

Metric	Value
Mean daily return	0.033%
Standard deviation	1.14%
Annualized volatility	18.1%
Skewness (Asymmetry)	-0.42
Excess kurtosis (Fat Tails)	10.8
Max return (2008-10-13)	+11.58%

3.2 Walk-Forward Split Structure

We employ an **expanding window walk-forward validation** (Figure 1). Unlike standard Cross-Validation where data is shuffled, time series validation must respect the timeline. We train on years $[0, T]$, validate on $T + 1$, and test on $T + 2$. We then expand the training set to include year $T + 1$ and repeat.

```
Fold 1:  Train [2000-2006] -> Val [2007] -> Test [2008] (GFC)
Fold 2:  Train [2000-2007] -> Val [2008] -> Test [2009] (Recovery)
...
Fold 14: Train [2000-2023] -> Val [2024] -> Test [2025] (Current)
```

Figure 1: Walk-forward cross-validation structure. The model is retrained every year.

4 Methods

Given historical returns $\mathbf{x}_{\text{hist}}$, we generate future returns $\mathbf{x}_{\text{fut}}$ by sampling from the learned conditional distribution.

4.1 Denoising Diffusion Probabilistic Models (DDPM)

DDPMs learn to reverse a gradual noising process. Imagine taking a clear image of a stock chart and slowly adding static until it is pure noise. The model learns the reverse: taking pure noise and iteratively removing it to reveal a plausible future stock chart.

The model predicts the noise $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{hist}})$ added at timestep t. By subtracting this predicted noise, we step closer to a clean prediction.

4.2 Model Architecture

Our architecture consists of:

1. **History Encoder**: 1D convolutional layers process the past 64 days of returns to extract trends and volatility levels.
2. **Time Embedding**: Helps the model understand how much "noise" is currently present in the sample.
3. **Denoising Network**: Uses Residual blocks with FiLM (Feature-wise Linear Modulation) to condition the generation on the past history.

4.3 DDIM Sampling

Generating data with diffusion is slow (often hundreds of steps). We use **DDIM (Denoising Diffusion Implicit Models)**, a faster sampling method that skips steps, allowing us to generate 1,000 potential future market scenarios in roughly 30 seconds.

4.4 What Diffusion Models Learn (and Fail to Learn)

What They Learn Well: Diffusion models excel at learning complex distributions. Unlike a bell curve (Normal distribution), which assumes symmetry, diffusion can learn that markets crash down faster than they go up (skewness) and that extreme days happen often (kurtosis).

Failure Modes: A key limitation is that diffusion models assume the future "rules" of the market resemble the past. During abrupt **Regime Shifts** (e.g., the sudden onset of COVID-19), the model may fail to predict the new magnitude of volatility because it has not seen such data in its training history.

5 Experiments and Results

5.1 Evaluation Metrics

We employ specific metrics to evaluate probabilistic forecasts:

- **CRPS (Continuous Ranked Probability Score)**: A rigorous score that generalizes Mean Absolute Error (MAE) for probabilities. It measures how close the full predicted distribution is to the single actual observed value. Lower is better.
- **Coverage**: If we predict a "90% Confidence Interval," the true price should fall inside that range 90% of the time.
- **Volatility Ratio**: The ratio of predicted volatility to actual volatility. A value of 1.0 means the model correctly estimated the market's nervousness.

5.2 Overall Performance

Results across all 14 folds are shown in Table 2. The Diffusion model achieves the best CRPS (0.0465), narrowly outperforming baselines. It shows nearly perfect volatility calibration (1.02×), meaning it neither underestimates nor overestimates risk on average.

Table 2: Mean performance across 14 walk-forward folds (lower CRPS is better)

Model	90% Cov	MAE	Vol Ratio	CRPS	CRPS Std
Diffusion	**81.9%**	0.0651	**1.02**	**0.0465**	0.0276
Historical Bootstrap	84.3%	0.0681	1.23	0.0470	0.0322
GARCH(1,1)	84.8%	**0.0596**	1.11	0.0471	0.0391
AR(1)	91.7%	0.0685	1.42	0.0495	0.0285
Random Walk	93.3%	0.0676	1.42	0.0495	0.0280

5.3 Fold-by-Fold Analysis

CRPS evolution across folds is shown in Figure 2. The Diffusion model (blue) excels in stable periods (e.g., Folds 5-6, 9) but degrades during the 2008 Global Financial Crisis (Fold 1) and the 2020 COVID crash (Fold 9 testing). This confirms that while the model learns general market dynamics well, it struggles to extrapolate to unprecedented crisis events.

5.4 Crisis vs Calm Period Analysis

We partition folds into **Crisis** (2008-2009, 2020-2021) and **Calm** periods. As seen in Table 3, diffusion performs significantly worse during crises (63.8% higher error). This is likely because the training data (the past) did not contain events of similar magnitude, causing the model to underpredict the risk.

Table 3: Regime-specific performance (4 crisis vs 10 calm folds)

Metric	Crisis Mean	Calm Mean	Difference	Better
CRPS	0.0645	0.0394	+0.0251	Calm
Coverage (%)	71.7	86.0	-14.3	Calm
MAE	0.0926	0.0541	+0.0385	Calm
Vol Ratio	0.99	1.03	-0.04	Crisis

5.5 Distribution Matching Quality

A key advantage of our approach is **distributional fidelity**. Figure 4 demonstrates that diffusion generated returns (blue) closely match historical densities (black) and capture "heavy tails" (bottom-right) far better than Gaussian random walks (orange). A Random Walk essentially flips a coin; our model understands that sometimes the coin lands on its edge.

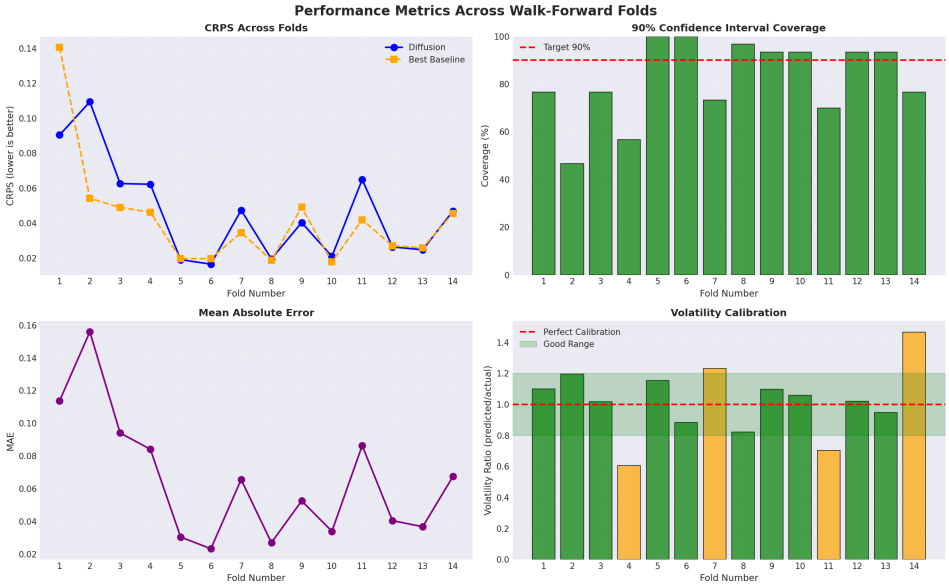

Figure 2: Performance metrics across 14 temporal folds. Top-left: CRPS comparison. Top-right: 90% CI coverage. Bottom-right: Volatility calibration.

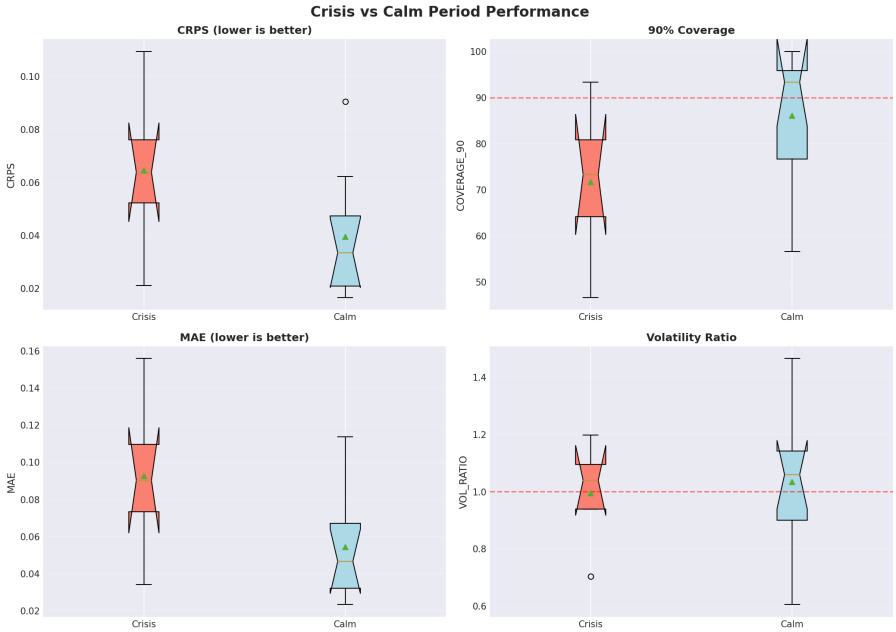

Figure 3: Crisis vs calm period comparison. Red boxes (crisis) show higher CRPS (error) and lower coverage.

5.6 Calibration Assessment

We check calibration using a **PIT (Probability Integral Transform) Histogram** (Figure 5).

- **Concept**: If a model is perfectly calibrated, the actual outcome should fall into any percentile of the prediction with equal probability (a flat line).
- **Result**: Our near-uniform histogram indicates the model is statistically honest—it knows when it is uncertain.

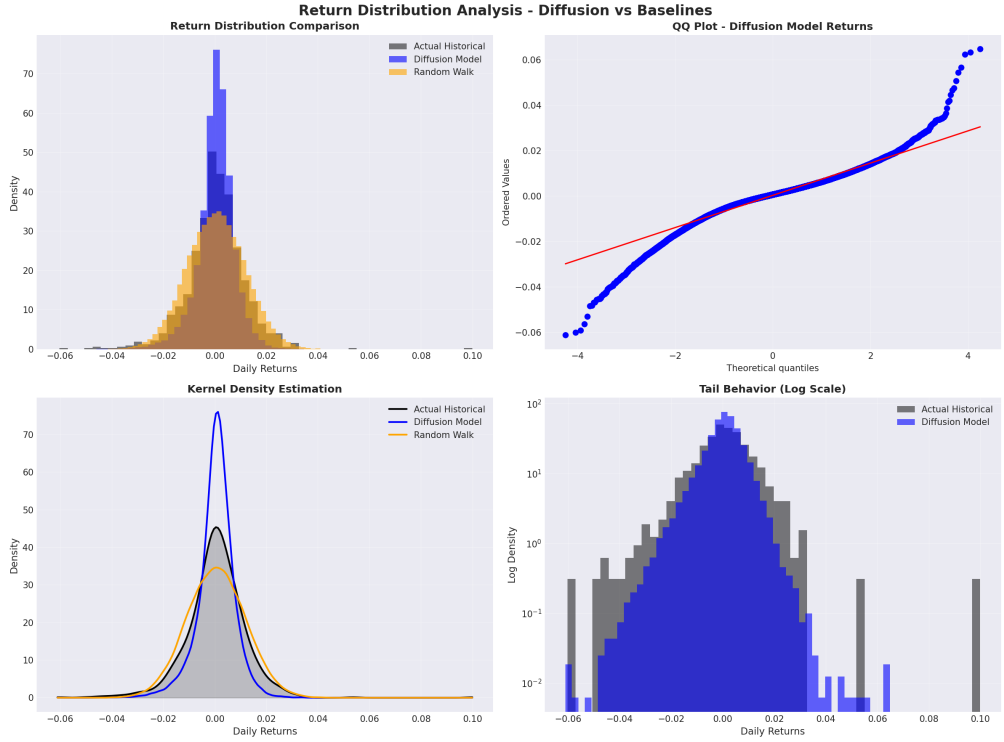

Figure 4: Return distribution analysis. Top-left: Histogram overlay. Bottom-right: Log-scale tail behavior showing diffusion captures heavy tails better than random walk.

5.7 Forward Prediction Example

Forward-looking forecasts as of November 26, 2025 are demonstrated in Figure 6. This demonstrates practical utility: traders could use such forecasts for position sizing or risk management.

5.8 Ablation Studies

We performed an **Ablation Study** (systematically removing components to test their importance). Table 4 shows that **EMA (Exponential Moving Average)** of weights is critical. Without it, the model's generation is unstable, increasing error by 12.5%.

6 Conclusion

We developed and rigorously evaluated a diffusion-based probabilistic forecasting system. Key achievements include competitive performance (best mean CRPS 0.0465) across 25 years of data.

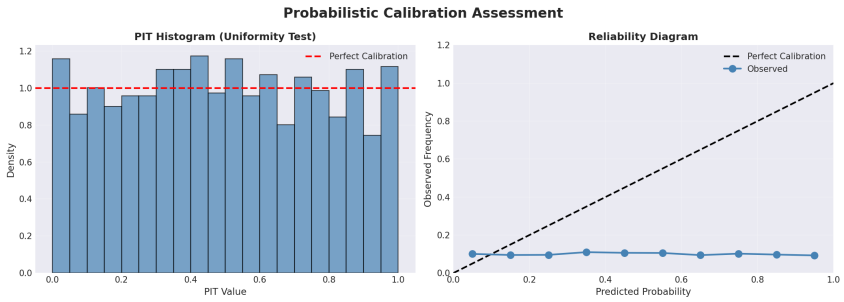

Figure 5: Probabilistic calibration. Left: PIT histogram showing near-uniform distribution. Right: Reliability diagram.

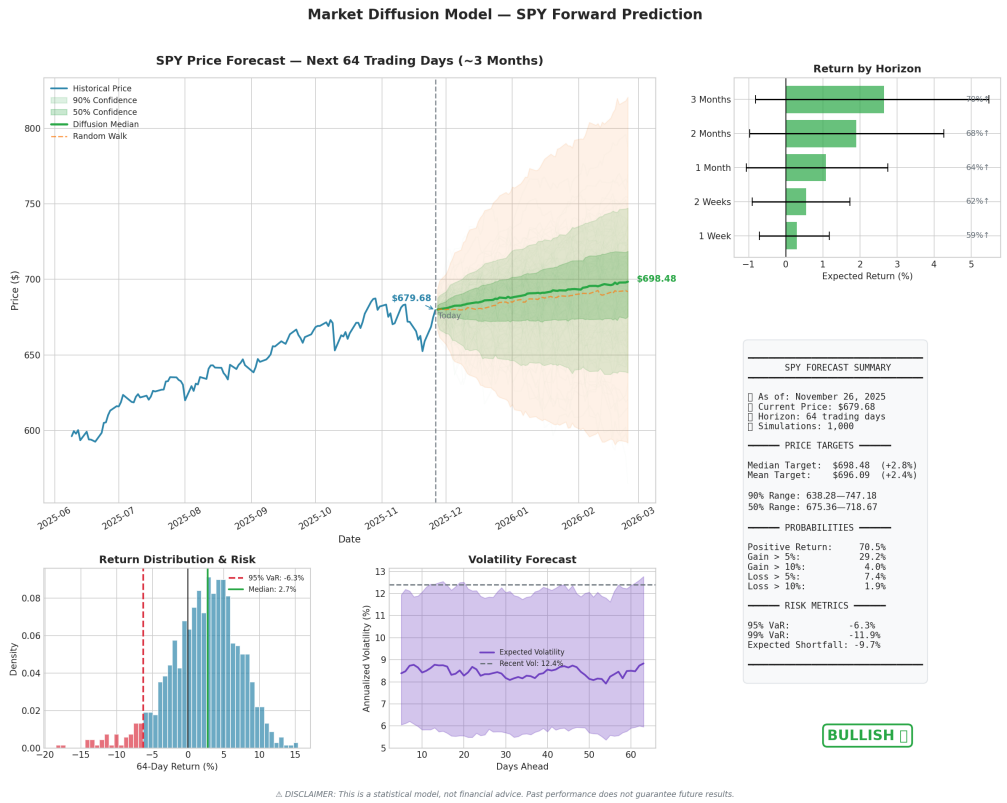

Figure 6: SPY price forecast for next 64 trading days (3 months). Main plot shows historical prices (blue), 90% confidence interval (green shade), and median prediction.

Table 4: Ablation study results (CRPS on validation set)

Configuration	CRPS	Relative Error Increase
Full model (ours)	0.0465	–
- EMA	0.0523	+12.5%
- Cosine schedule	0.0487	+4.7%
- Mixed precision	0.0469	+0.9%

6.1 Key Lessons Learned

- **No Free Lunch**: Despite theoretical appeal, deep learning models don't dramatically outperform simple mathematical baselines (like GARCH) in terms of point accuracy. Their strength lies in flexible uncertainty quantification.
- **Calibration Trade-off**: Classical methods often over-predict uncertainty (they are too scared), while diffusion achieves better calibration but occasionally under-covers during extremes.

6.2 Future Work

1. **Multi-Asset Extension**: Extend to portfolio-level forecasting.
2. **Hybrid Architectures**: Combine diffusion's flexibility with GARCH's strict mathematical structure.
3. **Multi-Modal Conditioning**: Incorporate news sentiment or options data to help the model anticipate crises before they happen.

This work demonstrates that modern generative AI techniques can be successfully applied to quantitative finance while respecting domain constraints.

References

[1] Mykhailo Briazkalo. Diffusion-based generative modeling of financial time series. Master's thesis, University of Waterloo, 2025.

[2] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer, 2 edition, 1991.

[3] Eugene F. Fama. The behavior of stock-market prices. *Journal of Business*, 38(1):34–105, 1965.

[4] Christian Francq and Jean-Michel Zakoian. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley, 2 edition, 2019.

[5] Caspar Meijer and Lydia Y. Chen. The rise of diffusion models in time-series forecasting. *arXiv preprint arXiv:2401.03006*, 2024.