

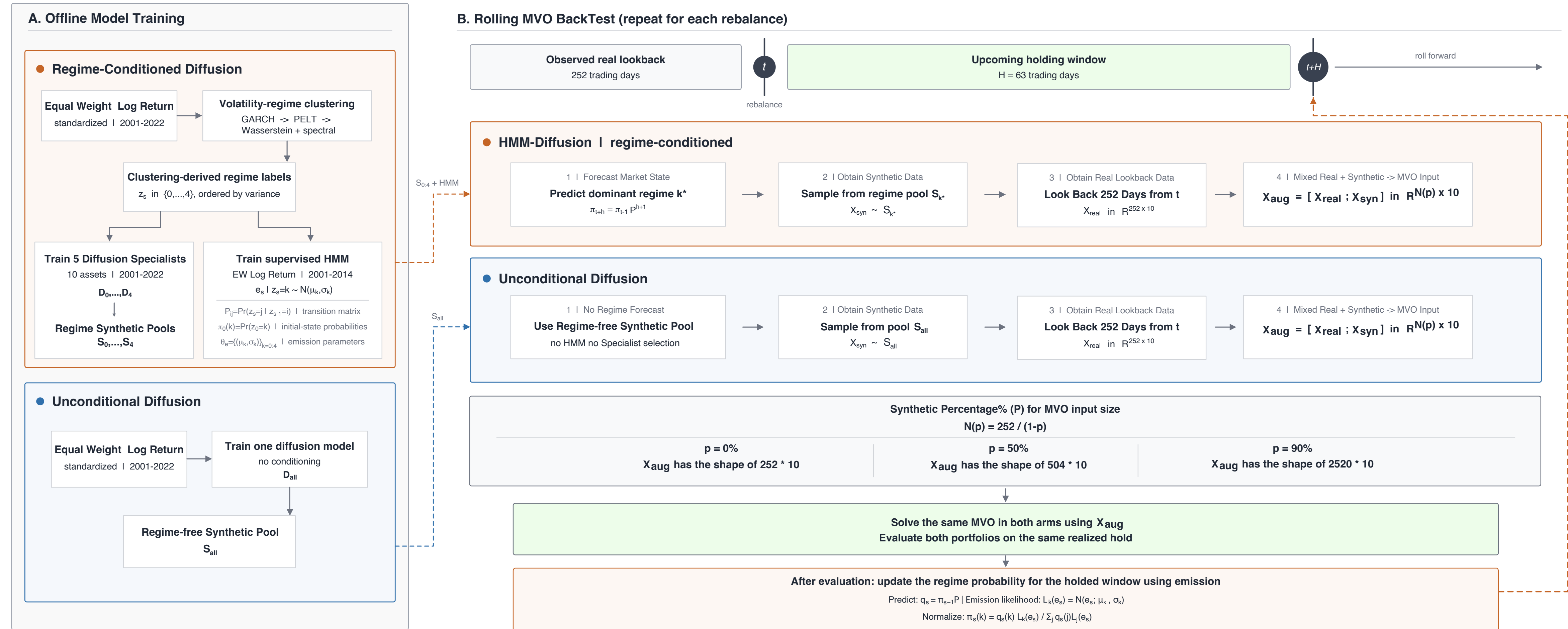
Introduction

Methodology Illustration

- Mean-variance optimization (MVO), volatility forecasting, and return prediction rely on historical returns, but limited observations and changing market regimes can make estimated relationships unstable.
- Diffusion models can expand the training sample with synthetic return paths. However, it remains unclear whether synthetic augmentation improves downstream performance relative to real-only training and whether regime conditioning provides additional value.
- We compare two approaches: **Unconditional Diffusion**, which learns one regime-free return distribution, and **HMM-Diffusion**, which trains regime-specific diffusion specialists and uses a supervised Hidden Markov Model to select the generator aligned with the forecasted holding-window regime.
- We vary the synthetic-data proportion while keeping the real lookback, asset universe, evaluation periods, and downstream models fixed.
- Downstream utility is evaluated using out-of-sample MVO Sharpe and Calmar ratios and the R-squared volatility forecasts across horizons. Forecasting features include lagged returns, multi-scale volatility, higher moments, and technical indicators.

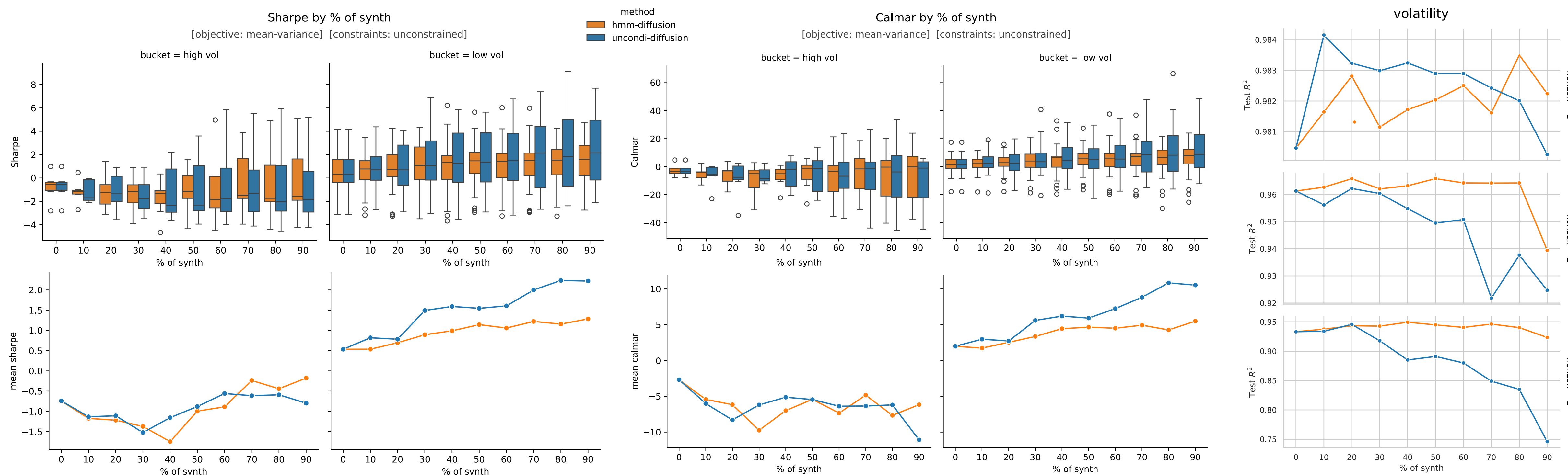
Research Questions:

- Does synthetic augmentation improve portfolio optimization and financial forecasting relative to real-only data?
- When does regime-conditioned generation outperform unconditional generation across downstream tasks and forecast horizons?
- How does the effect of synthetic augmentation on portfolio performance differ between high- and low-volatility market windows?



Results

Conclusions



- Portfolio Optimization:** In low-volatility windows, increasing the synthetic proportion produces an overall upward trend in both Sharpe and Calmar ratios. Unconditional Diffusion consistently outperforms HMM-Diffusion, indicating that broader regime-free samples are more effective for MVO under stable market conditions. Neither method produces a consistent improvement in high-volatility windows.
- Volatility Forecasting:** Both methods can improve short-horizon volatility forecasts at selected synthetic proportions. At longer horizons, HMM-Diffusion maintains substantially more stable R-squared values as the synthetic share increases, whereas Unconditional Diffusion deteriorates at higher augmentation levels. Regime conditioning therefore improves the robustness of synthetic augmentation for longer-horizon volatility forecasting.
- Overall Finding:** The value of synthetic financial data is task-dependent. Unconditional augmentation is more effective for low-volatility portfolio construction, while regime-conditioned augmentation better preserves predictive information for longer-horizon volatility forecasting. Synthetic-data design should therefore be matched to the downstream objective rather than treated as universally beneficial.