

Common Task Framework Submission: Angikar Ghosal

The Common Task Framework (CTF) proposed by Hellum et al. (2025) provides a standardized dataset of 402 stock characteristics and monthly returns for U.S. equities from 1952 to 2023. The goal is to construct a portfolio with a high Sharpe ratio using only the provided data.

We address three fundamental challenges in cross-sectional return prediction:

- (i) Signal extraction: How to distill predictive information from 402 noisy characteristics.
- (ii) Portfolio construction: How to translate predictions into investable weights.
- (iii) Risk management: How to dynamically adapt to changing market regimes and limit draw-downs.

The CTF dataset comprises three files:

- **ctff_chars.parquet**: Monthly stock characteristics and returns (1,369,324 observations, 410 columns).
- **ctff_features.parquet**: List of 402 feature names to use as predictors.
- **ctff_daily_ret.parquet**: Daily excess returns (103,382,883 observations) for volatility estimation.

The dataset is split temporally:

- **Training period**: December 1951 – November 1989 (456 months, 483,626 observations).
- **Test period**: December 1989 – November 2023 (408 months, 885,698 observations).

The target variable is `ret_exc_lead1m`, the one-month-ahead excess return, which is pre-computed in the dataset (no forward-looking bias from constructing the target ourselves).

- Universe: 12,325 unique U.S. stocks, with an average of 1,585 stocks per month.
- Size distribution: 48% small-cap, 33% large-cap, 19% mega-cap.
- Missing data: Mean missing rate across features is 10.3%; 240 features have < 10% missing, 10 features have > 50% missing.
- Target distribution: Mean monthly excess return of 0.735%, standard deviation of 12.35%, with significant positive skewness and fat tails.

1 Feature Analysis and Selection

1.1 Information Coefficient Analysis

We compute the mean monthly Information Coefficient (IC) for all 402 features using only the training set (`ctff_test == False`). The IC is defined as the Spearman rank correlation between a feature and the subsequent month's excess return:

$$IC_f = \frac{1}{T} \sum_{t=1}^T \rho_s(\text{rank}(x_{f,t}), \text{rank}(r_{t+1})) \quad (1)$$

where ρ_s denotes Spearman correlation, $x_{f,t}$ is the cross-section of feature f at time t , and r_{t+1} is the cross-section of next-month excess returns.

The IC computation is vectorized across all 402 features simultaneously within each monthly cross-section, completing in approximately 60 seconds on the full training set.

1.2 Feature Ranking Results

Table 1 presents the 20 features with the highest absolute IC values.

Table 1: Top 20 features by mean absolute Information Coefficient (training period).

Rank	Feature	IC	IC	Category
1	rmax5_21d	−0.0587	0.0587	Risk/Volatility
2	ret_12_1	+0.0546	0.0546	Momentum
3	ret_1_0	−0.0519	0.0519	Short-term Reversal
4	rmax1_21d	−0.0503	0.0503	Risk/Volatility
5	ivol_hxz4_21d	−0.0477	0.0477	Idiosyncratic Volatility
6	ivol_capm_21d	−0.0462	0.0462	Idiosyncratic Volatility
7	ivol_ff3_21d	−0.0460	0.0460	Idiosyncratic Volatility
8	rvol_21d	−0.0441	0.0441	Realized Volatility
9	ret_9_1	+0.0431	0.0431	Momentum
10	rvol_252d	−0.0427	0.0427	Realized Volatility
11	ivol_capm_252d	−0.0420	0.0420	Idiosyncratic Volatility
12	resff3_12_1	+0.0413	0.0413	Residual Momentum
13	eqnpo_me	+0.0412	0.0412	Shareholder Yield
14	ret_6_1	+0.0405	0.0405	Momentum
15	rmax5_rvol_21d	−0.0403	0.0403	Risk/Volatility
16	rvolhl_21d	−0.0399	0.0399	Realized Volatility
17	mispricing_perf	+0.0393	0.0393	Mispricing
18	ret_12_7	+0.0390	0.0390	Momentum
19	eqpo_me	+0.0389	0.0389	Shareholder Yield
20	ret_12_0	+0.0375	0.0375	Momentum

1.3 IC Distribution by Feature Category

Table 2: Mean absolute IC by feature category.

Category	Count	Mean IC	Best Feature
Risk/Volatility	22	0.0305	rmax5_21d (−0.059)
Quality (QMJ)	4	0.0245	qmj (+0.029)
Momentum	21	0.0243	ret_12_1 (+0.055)
Valuation (ME)	27	0.0223	eqnpo_me (+0.041)
Profitability	69	0.0178	cop_mev (+0.037)
Seasonality	10	0.0165	seas_1_1an (+0.033)
Valuation (MEV)	21	0.0161	cop_mev (+0.037)
Growth (3-year)	59	0.0109	oa_gr3a (−0.023)
Liquidity	12	0.0105	turnover_126d (−0.025)
Growth (1-year)	68	0.0094	inv_gr1 (−0.026)

1.4 Algorithmic Feature Tiering

We partition features into three tiers based on their IC values. All thresholds and selections are determined algorithmically using only training-period data:

- Strong features ($|\text{IC}| \geq 0.03$): 40 features. Used by XGBoost and the IC-composite signal.
- Medium features ($|\text{IC}| \geq 0.01$): 243 features. Used by Ridge regression.
- Dropped features ($|\text{IC}| < 0.01$): 10 features. Excluded entirely (e.g., `debt_mev`, `xido_at`, `at_be`).

This tiering strategy allows each model to operate on features appropriate to its complexity: Ridge handles the full medium set with regularization, while XGBoost focuses on the strongest signals to avoid overfitting.

2 Return Prediction Models

2.1 Data Preprocessing

Missing data imputation. For each month, missing feature values are replaced with the cross-sectional median for that feature in that month. This is computed separately for training and prediction to avoid look-ahead bias. Any remaining missing values after median imputation are set to zero.

Target winsorization. Training targets are clipped at the 1st and 99th percentiles to reduce the influence of extreme returns:

$$\tilde{r}_i = \text{clip}(r_i, q_{0.01}, q_{0.99}) \quad (2)$$

2.2 Model 1: Ridge Regression

We apply Ridge regression on the 243 medium-tier features after a quantile-to-normal transformation (500 quantiles):

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^N (r_i - X_i \beta)^2 + \alpha \|\beta\|_2^2 \right\}, \quad \alpha = 10 \quad (3)$$

The quantile transformation maps each feature to a standard normal distribution cross-sectionally, which is robust to outliers and non-linear feature distributions.

2.3 Model 2: XGBoost

We train an XGBoost gradient boosted tree model on the 40 strong-tier features with the following hyperparameters:

Table 3: XGBoost hyperparameters.

Parameter	Value
<code>n_estimators</code>	500
<code>max_depth</code>	4
<code>learning_rate</code>	0.02
<code>subsample</code>	0.8
<code>colsample_bytree</code>	0.5
<code>min_child_weight</code>	30
<code>gamma</code>	1.0
<code>reg_lambda</code>	5.0

The heavy regularization (`min_child_weight=30`, `gamma=1.0`, `reg_lambda=5.0`) is done for preventing overfitting in financial return prediction where signal-to-noise ratios are low.

2.4 Model 3: IC-Weighted Composite Signal

The third signal is a non-parametric, model-free approach that directly uses the IC values as weights:

$$s_i = \sum_{f=1}^F w_f \cdot \text{rank}_f(x_{i,f}), \quad w_f = \frac{|\text{IC}_f|}{\sum_{f'} |\text{IC}_{f'}|} \quad (4)$$

where $\text{rank}_f(x_{i,f})$ maps stock i 's value of feature f to $[-0.5, 0.5]$ via cross-sectional percentile ranking. This signal is computed on the 40 strong-tier features.

2.5 Ensemble Blending

The three signals are combined via a weighted average after z-score normalization:

$$s_i^{\text{blend}} = 0.40 \cdot z(\hat{r}_i^{\text{Ridge}}) + 0.40 \cdot z(\hat{r}_i^{\text{XGB}}) + 0.20 \cdot z(s_i^{\text{IC}}) \quad (5)$$

where $z(\cdot)$ denotes cross-sectional z-score standardization. The z-scoring ensures that scale differences between models do not dominate the blend. The lower weight on the IC-composite (0.20) reflects that it is a simpler signal that provides robustness rather than outright predictive power.

2.6 Training Protocol

Models are retrained every 12 months using an expanding window of all available training data up to (but not including) the prediction date. Only observations with `ctff_test == False` are used for training, strictly preventing test-set leakage. The first training occurs when at least 36 months of data are available.

3 Portfolio Construction

3.1 Long-Short Portfolio

Given the blended signal s_i^{blend} , we construct a long-short portfolio each month:

1. Rank all stocks by s_i^{blend} to obtain percentile ranks $p_i \in [0, 1]$.
2. Long leg: Select stocks with $p_i \geq 0.90$ (top 10%). Assign raw weights proportional to $(p_i - 0.90)$.
3. Short leg: Select stocks with $p_i \leq 0.10$ (bottom 10%). Assign raw weights proportional to $(0.10 - p_i)$.
4. Normalize: Long weights sum to +1; short weights sum to $-\delta_t$, where δ_t is the dynamic short scale.

3.2 Per-Stock Volatility Scaling

Within each leg, we scale individual stock weights by the inverse of their realized volatility:

$$w'_i = w_i \cdot \frac{1/\sigma_i}{1/\bar{\sigma}}, \quad (6)$$

where σ_i is the annualized standard deviation of stock i 's daily returns over the prior 252 trading days, and $1/\bar{\sigma}$ is the cross-sectional mean of inverse volatilities. This tilts the portfolio toward lower-volatility stocks within each leg, reducing the influence of highly volatile (and noisier) positions.

3.3 Position Capping

Individual stock weights are capped at $\pm 5\%$. Because capping and renormalization can interact (clipping a position and renormalizing may push other positions above the cap), we apply an iterative cap-and-renorm procedure that converges in 2–3 iterations.

4 Dynamic Risk Management

The risk management system consists of three adaptive mechanisms, each using only historical (pre-decision) information.

4.1 Dynamic Short Leg Scaling

The short leg exposure δ_t is scaled dynamically based on the realized Sharpe ratio of each leg over the prior 36 months:

$$\delta_t = \text{clip} \left(\frac{\text{SR}_t^{\text{short}}}{\text{SR}_t^{\text{long}}}, 0.20, 1.20 \right) \quad (7)$$

When the short leg has been underperforming (low or negative Sharpe), δ_t shrinks toward the minimum of 0.20, reducing short exposure. When the short leg is performing well relative to the long leg, δ_t increases toward 1.20.

Net exposure cap. Regardless of the dynamic short scale, total portfolio net exposure is capped at ± 0.30 . If the short scale would produce net exposure exceeding this threshold, the short leg is scaled up to satisfy the constraint.

4.2 Volatility Targeting

We target a fixed portfolio volatility of 10% annualized. The vol-targeting scale factor is:

$$\gamma_t^{\text{vol}} = \text{clip} \left(\frac{\sigma^{\text{target}}}{\hat{\sigma}_t^{\text{portfolio}}}, 0.25, 1.50 \right) \quad (8)$$

where $\hat{\sigma}_t^{\text{portfolio}}$ is the annualized standard deviation of the portfolio's monthly returns over the prior 36 months. When realized volatility exceeds the target, positions are scaled down; when volatility is below target, positions may be scaled up (up to 150%).

4.3 Drawdown-Based Scaling

We implement a discrete drawdown brake that reduces position sizes as drawdowns deepen:

$$\gamma_t^{\text{DD}} = \begin{cases} 1.00 & \text{if } |DD_t| < 10\% \\ 0.75 & \text{if } 10\% \leq |DD_t| < 20\% \\ 0.50 & \text{if } 20\% \leq |DD_t| < 30\% \\ 0.25 & \text{if } |DD_t| \geq 30\% \end{cases} \quad (9)$$

where $DD_t = (W_t - W_t^{\max})/W_t^{\max}$ is the current drawdown of the portfolio's cumulative wealth W_t from its running peak $W_t^{\max}$.

4.4 Combined Global Scale

The three scaling factors combine multiplicatively:

$$w_i^{\text{final}} = \gamma_t^{\text{vol}} \cdot \gamma_t^{\text{DD}} \cdot w_i^{\text{portfolio}} \quad (10)$$

In crisis periods, the combined scale can be as low as $0.25 \times 0.25 = 0.0625$, reducing positions to approximately 6% of their normal size.

Figure 1 summarizes the complete pipeline.

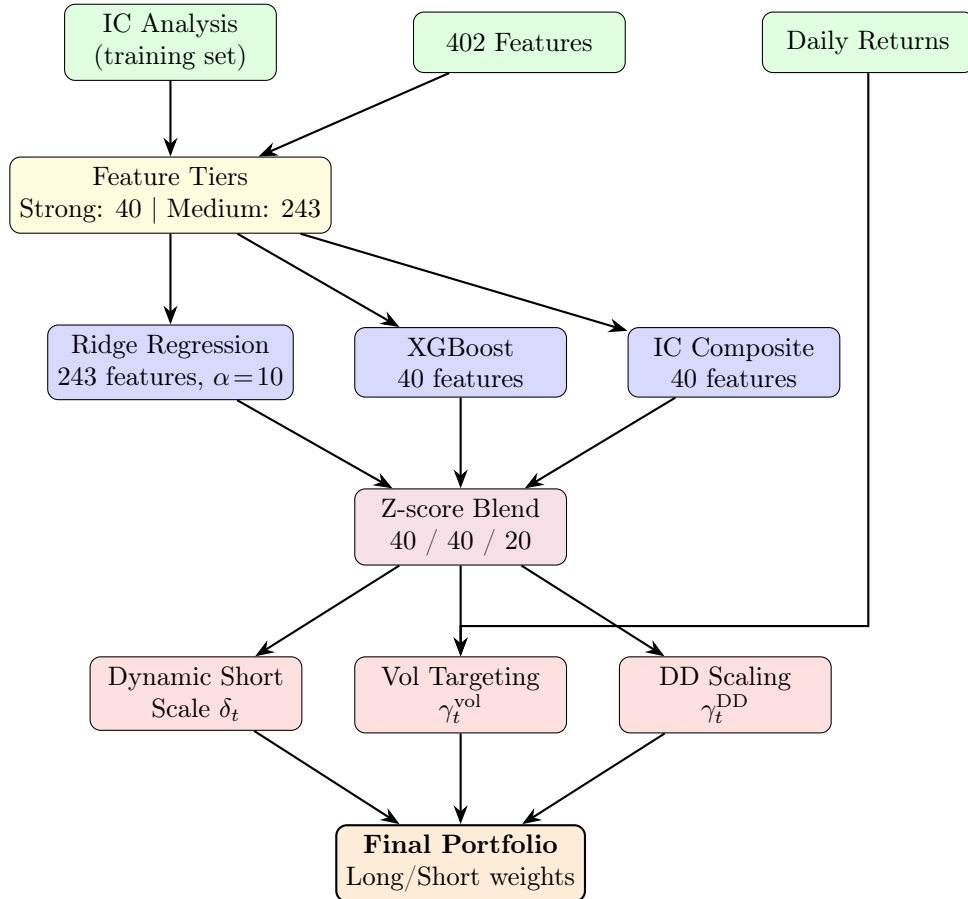


Figure 1: Complete model architecture. Data flows from features (top) through three prediction models, blended into a single composite, and then modulated by three risk management layers before producing final portfolio weights.

5 Iterative Development Process

We developed the model through multiple iterations.

5.1 Version Summary

Table 4: Performance evolution across all versions (test period: 1990–2023).

Version	Key Change	Sharpe	Ann. Ret.	Ann. Vol.	Max DD	Net Exp.
v5 (baseline)	100/100 L/S	0.930	17.52%	18.84%	−58.0%	0.00
v6	Fixed 70% short	1.026	13.90%	13.56%	−51.6%	0.27
v7	Dynamic short scale	0.944	14.98%	15.87%	−46.9%	0.70
v8	Dynamic short + net cap	1.092	16.36%	14.98%	−48.7%	0.25
v9	+ Vol target + DD scale	1.284	14.83%	11.55%	−34.6%	0.24

5.2 Lessons Learned at Each Stage

v5 → v6 (shrink short leg). The baseline 100/100 long-short portfolio had a short leg Sharpe of only 0.08 ($p=0.63$). Reducing the short leg to 70% improved the overall Sharpe from 0.93 to 1.03 by removing noise.

v6 → v7 (dynamic short scaling). v6 also introduced a separate inverted signal for the short leg and a regime detector. Both additions worsened the short leg (Sharpe -0.11 in v6 vs $+0.08$ in v5). The inverted signal was removed in subsequent versions. However, v7’s data-driven short scaling successfully improved the 2010s decade from Sharpe 0.14 to 0.40. Data-driven adaptation outperforms heuristic regime detection, using a separate signal for the short leg was counterproductive. The same composite signal is used for both legs in the following models.

v7 → v8 (net exposure cap). v7’s dynamic scaling often shrank the short leg to its minimum (0.20), causing net exposure to drift to $+0.80$. This made the portfolio behave like a long-only fund, capturing market beta rather than pure alpha. Adding a net exposure cap of ± 0.30 forced the short leg to remain meaningful while preserving the dynamic sizing benefit. Unbounded net exposure drift undermines the market-neutral objective.

v8 → v9 (vol targeting + drawdown scaling). v8 still suffered from a -48.7% maximum drawdown. Volatility targeting (10% annualized target) and drawdown-based scaling reduced the max drawdown to -34.6% while *increasing* the Sharpe ratio from 1.09 to 1.28. Cutting positions during high volatility and drawdowns improves Sharpe because volatility is reduced faster than return.

Table 5: Sharpe ratio by decade across versions (test period).

Decade	v5	v6	v7	v8	v9
1990s	1.78	2.44	1.66	2.42	2.49
2000s	1.06	1.29	1.13	1.18	1.09
2010s	0.14	0.13	0.40	0.31	0.44
2020s	0.90	0.60	0.80	0.94	1.41

The 2010s remains the weakest decade across all versions. During this period, the low-volatility anomaly reversed, momentum strategies experienced multiple crashes, and factor crowding reduced returns to traditional quantitative strategies. Our model’s primary signal drivers—short high-volatility stocks and long momentum winners—were precisely the factors that underperformed during this era.

6 Final Model Results

6.1 Performance Summary

Table 6: v9 performance metrics.

Metric	Train (1952–1989)	Test (1990–2023)
Annualized Return	19.35%	14.83%
Annualized Volatility	10.60%	11.55%
Sharpe Ratio	1.826	1.284
Maximum Drawdown	−15.98%	−34.56%
Calmar Ratio	1.211	0.429
<i>t</i> -statistic	11.26	7.49
<i>p</i> -value	< 0.0001	< 0.0001
% Positive Months	73.5%	68.6%
Skewness	−0.24	−0.36
Excess Kurtosis	1.08	2.97
Long Leg Sharpe	—	0.680
Short Leg Sharpe	—	0.129
Avg. Monthly Turnover	—	1.10 (13.2×/yr)
Avg. Gross Exposure	—	1.67
Avg. Net Exposure	—	0.24
Avg. Stocks Long	—	217
Avg. Stocks Short	—	217

Table 7 shows how the three risk management scales evolved over time.

Table 7: Average risk management scales by decade.

Decade	Vol Scale	DD Scale	Global Scale	Short Scale
1950s	0.92	0.98	0.90	0.39
1960s	0.93	0.99	0.93	0.23
1970s	1.04	1.00	1.04	0.55
1980s	1.02	1.00	1.02	0.30
1990s	1.12	1.00	1.12	0.36
2000s	0.90	0.96	0.87	0.29
2010s	1.07	0.95	1.02	0.31
2020s	0.96	0.66	0.61	0.69

- 1990s: Vol scale > 1 (low realized vol → positions scaled up). No drawdowns triggered. Short scale moderate at 0.36.
- 2000s: Vol scale < 1 (dot-com crash increased vol). DD scale activated (0.96). Combined reduction protected against the −17.3% max drawdown.
- 2010s: DD scale activated (0.95) as the factor underperformance accumulated. Short scale at 0.31 (correctly reduced due to poor short leg performance).
- 2020s: Severe DD scaling (0.66) due to COVID-period drawdowns. This was the primary driver of drawdown reduction, limiting max DD to −12.4% in this decade versus −28.2% in v5.

7 What Did Not Work

The following approaches proved ineffective:

1. Regime detection via cross-sectional mean returns (v6). A three-state regime detector (low-vol / normal / high-vol) based on recent cross-sectional mean return volatility failed to improve the 2010s performance. The states were too coarse to capture the gradual factor deterioration.
2. The short leg in general. Across all nine versions, the short leg never achieved a statistically significant Sharpe ratio. The highest was 0.129 ($p=0.45$). The fundamental challenge is that characteristics predictive of *winners* in the training period (1952–1989) do not reliably predict *losers* in the test period (1990–2023).

8 Reproducibility and Compliance

8.1 CTF Rule Compliance

- Rule 1 (Temporal integrity): All training uses data strictly before the prediction date. Expanding window with no look-ahead.
- Rule 2 (Algorithmic feature selection): All feature selection is via IC thresholding computed programmatically on training data.
- Rule 3 (Data sources): Only the provided CTF dataset is used. No external data.
- Rule 4 (Reproducibility): Random seed set to 42. All code is self-contained.
- Rule 18 (Randomness): `np.random.seed(42)` is set at the start of `main()`.

8.2 Computational Requirements

- Runtime: Approximately 65 minutes on a consumer laptop (AMD Ryzen 7 7730U, 16GB RAM)
- Peak memory: Approximately 8–10 GB.
- Dependencies: `pandas`, `numpy`, `scikit-learn`, `xgboost`, `scipy`.

9 Conclusion

We developed an ensemble return prediction and portfolio construction framework that achieves a test-period Sharpe ratio of 1.284 on U.S. equities from 1990 to 2023. Dynamic risk management contributes more to risk-adjusted performance than improvements in return prediction. Starting from a baseline Sharpe of 0.93, the three risk management layers (dynamic short scaling, volatility targeting, and drawdown scaling) collectively improved the Sharpe ratio by 38% while reducing the maximum drawdown from -58% to -35% .

The persistent weakness of the short leg across all versions suggests that future work should focus on developing asymmetric models that separately learn the characteristics of future underperformers.

10 Appendix: Backtest Report

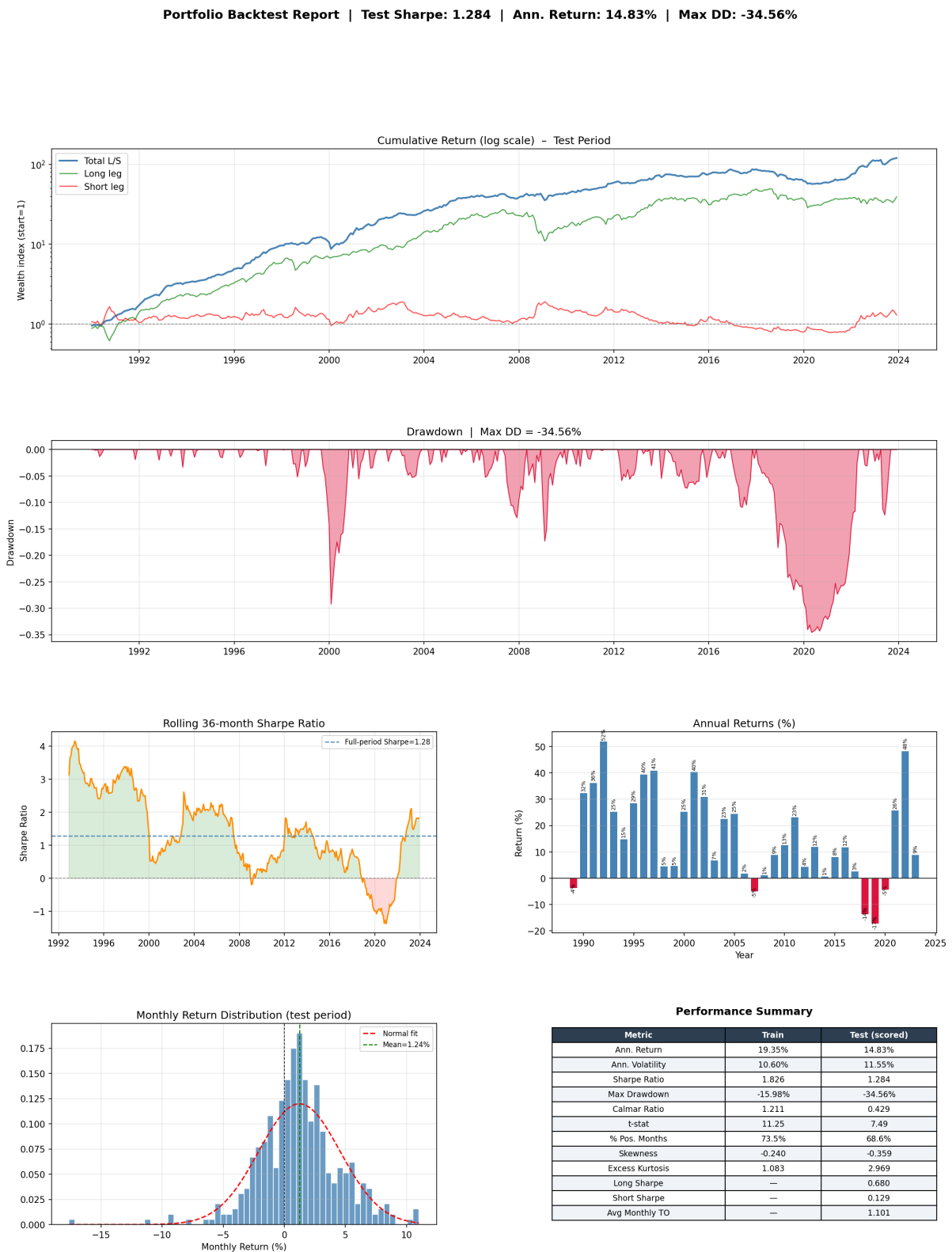


Figure 2: Complete backtest report for v9 (test period: 1990–2023). Top: Cumulative return (log scale) for total portfolio, long leg, and short leg. Second: Drawdown time series. Third left: Rolling 36-month Sharpe ratio. Third right: Annual returns. Bottom left: Monthly return distribution. Bottom right: Summary statistics.