

1 · Introduction

Question. Do characteristic *histories* carry information about conditional factor exposures beyond the latest cross-section? How to estimate highly parameterized attention-based asset pricing models in the low signal-to-noise ratio setting?

This paper. SACT (Saturated Attention Conditional Transformer) has these contributions: (i) saturated attention; (ii) a conditional factor model built on top of it; (iii) shows the economic value of attention architecture and characteristic paths; (iv) shows that ML models is complementary to conventional factor timing.

2 · Model Summary

Input — characteristic history

$$\mathbf{Z}_{i,t-1} = [\mathbf{z}_{i,t-L}, \dots, \mathbf{z}_{i,t-1}]^T \in \mathbb{R}^{L \times P}$$

Conditional factor structure

$$\mathbf{r}_{i,t} = \boldsymbol{\beta}(\mathbf{Z}_{i,t-1})^T \mathbf{f}_t + \varepsilon_{i,t}$$

Tradable latent factors

$$\mathbf{f}_t = \mathbf{W}_F \cdot (\mathbf{Z}_t^T \mathbf{Z}_t + \lambda_t \mathbf{I})^{-1} \mathbf{Z}_t^T \mathbf{r}_t$$

$\boldsymbol{\beta}(\cdot): \mathbb{R}^{L \times P} \rightarrow \mathbb{R}^K$ is the transformer mapping; its final causal token becomes the loading vector. $\mathbf{f}_t$ is a learned latent factor portfolio by rotating Fama–MacBeth characteristic-managed portfolios, so every latent factor maps back to stock holdings.

Estimated jointly on pricing MSE + L_2 , with early stopping and 10 ensemble models over different random initialization.

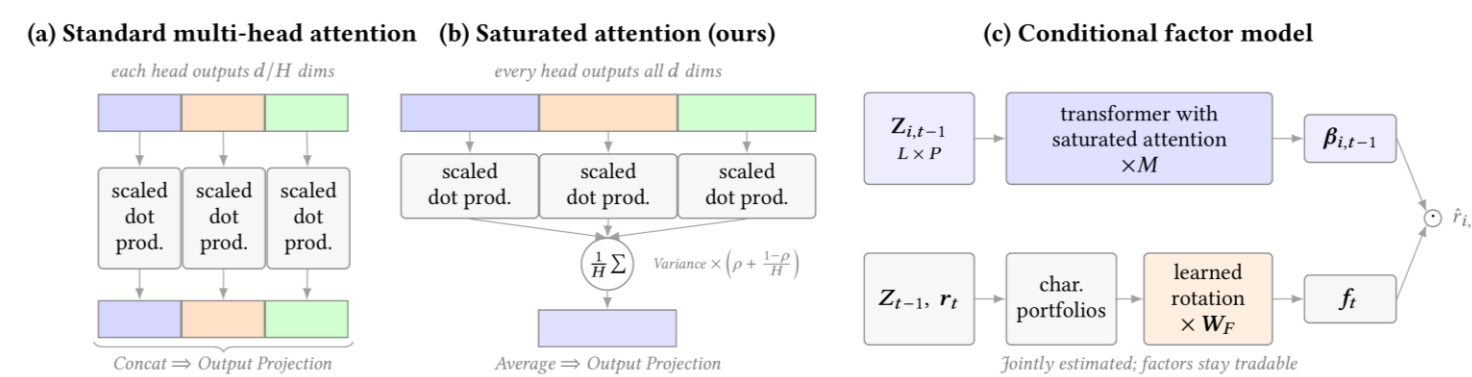
3 · Data & Experiment

Data. CRSP monthly U.S. returns, 1975–2024; strictly out-of-sample 2000–2024. 153 characteristics (Jensen–Kelly–Pedersen, 13 themes), rank-normalized to $[-0.5, 0.5]$, lagged 1/4/6 months for reporting delays. >17,000 firms, $\approx 4,000$ per month; no sector, price or missingness filters.

Protocol. Rolling 10y train / 5y validation / 1y test; annual refit, monthly rebalance.

Portfolio. Monthly tangency portfolio over the K latent factors, mapped back to the **largest 1,000 stocks** by market cap, $\|\mathbf{w}\|_1 = 2$ ($\approx 100/100$ long–short) for every model.

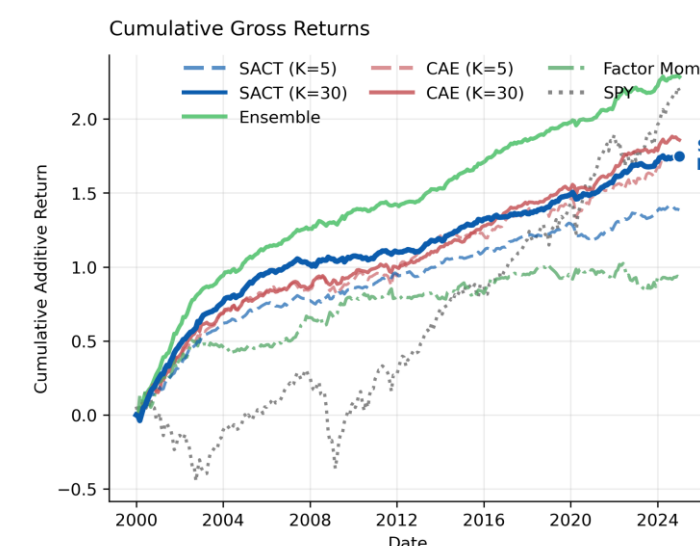
4 · Saturated Attention



- (a) Standard multi-head attention projects into H equal dimension subspaces and recombines them through a learned output projection, so the heads estimate complementary pieces rather than replicates. Under the common fixed-total-width convention, the design requires $H \mid d$.
- (b) Saturated attention instead keeps a d -dimensional output in every head and averages the resulting summaries, which contracts estimation variance.
- (c) The operator parameterizes path-dependent loadings $\boldsymbol{\beta}$, while a linear factor branch builds latent factors $\mathbf{f}_t$ from characteristic-managed portfolios so that every factor maps back to stocks.

5 · Empirical Results

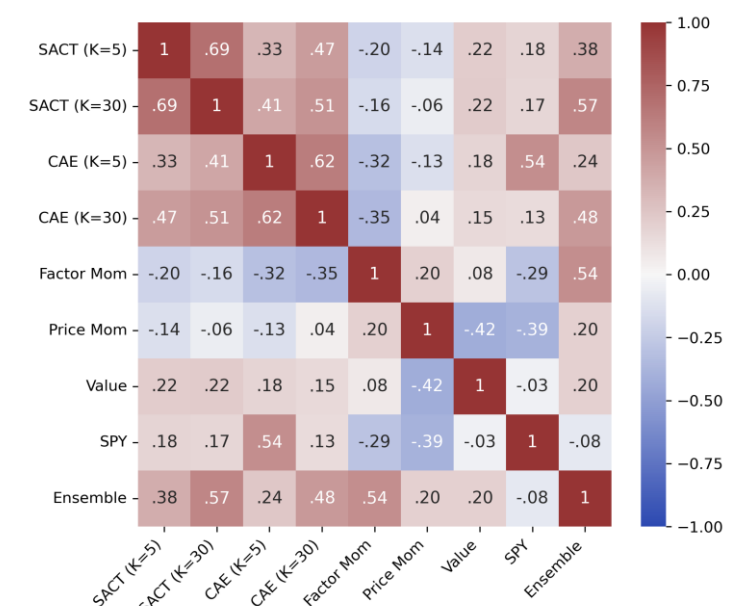
Gross Performance	Sharpe	Sortino	Ann. Return	Max DD	β vs SPY
SACT (K = 5)	1.42	2.29	5.5%	11.0%	0.045
SACT (K = 10)	1.53	2.84	5.8%	6.9%	0.035
SACT (K = 20)	1.80	3.10	6.8%	7.1%	0.036
SACT (K = 30)	1.80	3.48	7.0%	4.9%	0.044
SACT (K = 40)	1.71	2.93	6.5%	6.4%	0.031
CAE (K = 5)	1.12	1.54	6.9%	13.0%	0.221
CAE (K = 10)	1.28	1.84	6.8%	10.3%	0.106
CAE (K = 20)	1.30	1.72	6.4%	10.0%	0.073
CAE (K = 30)	1.57	2.44	7.4%	7.0%	0.042
CAE (K = 40)	1.44	2.39	6.5%	9.3%	0.023
Factor Momentum	0.45	0.65	3.7%	18.7%	-0.156
SPY	0.51	0.72	7.8%	50.8%	1.000
Ensemble (SACT + CAE + Factor Mom.)	2.11	4.18	9.4%	3.6%	-0.017



Net of 10bps Cost	Sharpe	Sortino	Turnover
SACT (K=30)	1.26	2.30	1.77
CAE (K=30)	0.99	1.53	2.31
SACT (K=5)	0.95	1.49	1.53
CAE (K=5)	0.75	1.03	1.80
Factor Momentum	0.00	0.00	2.81
Ensemble	1.49	2.90	2.18

Ablations. Saturated vs concat-head at $H=3$: Sharpe 1.73 / 1.62; post-2020 1.50 / 0.93. Context length $L=1$ vs $L=120$: post-2020 Sharpe 1.02 $\rightarrow$ 1.37;

6 · Strategy Correlation

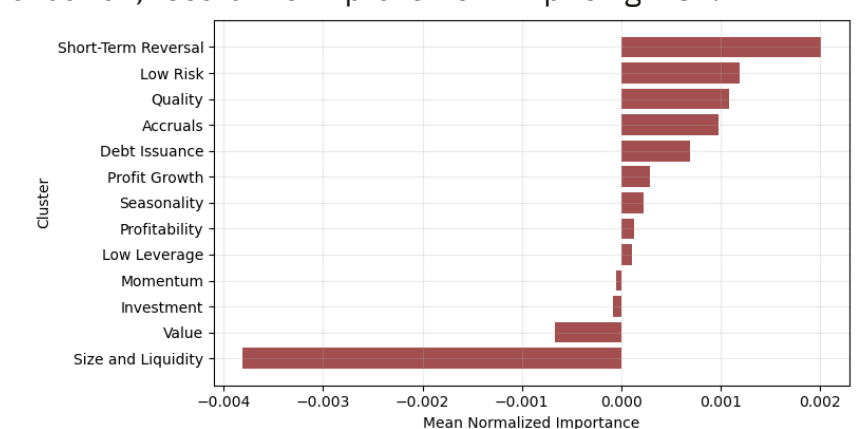


SACT and CAE consume identical characteristics yet correlate between **0.33** and **0.51** — far from redundant.

Factor momentum is **negatively** correlated with both ML models, so SACT is no repackaged factor-timing rule. Factor momentum earns its ensemble weight through covariance, not stand-alone return (gross Sharpe 0.45).

7 · Feature Importance

Perturbation importance: zero-fill one characteristic in validation, record the improvement in pricing MSE.



8 · Takeaways

- 1 SACT beats the CAE at every K , gross and net, with lower turnover (1.77 vs 2.31).
- 2 Characteristic paths pay off where point-in-time signals weaken: post-2020 Sharpe 1.02 ($L=1$) $\rightarrow$ **1.37** ($L=120$).
- 3 Saturated attention makes the sequence model work in low SNR data — Sharpe 1.62 $\rightarrow$ **1.73**, Sortino 2.66 $\rightarrow$ **3.58**.
- 4 Honest bound: prior to 2020, majority of the gap to CAE is architectural, not temporal.