# Kronos: A Foundation Model for the Language of Financial Markets

(https://arxiv.org/pdf/2508.02739)

**Abstract:** The success of large-scale pre-training paradigm, exemplified by Large Language Models (LLMs), has inspired the development of Time Series Foundation Models (TSFMs). However, their application to financial candlestick (K-line) data remains limited, often underperforming non-pre-trained architectures. Moreover, existing TSFMs often overlook crucial downstream tasks such as volatility prediction and synthetic data generation. To address these limitations, we propose Kronos, a unified, scalable pre-training framework tailored to financial K-line modeling. Kronos introduces a specialized tokenizer that discretizes continuous market information into token sequences, preserving both price dynamics and trade activity patterns. We pre-train Kronos using an autoregressive objective on a massive, multi-market corpus of over 12 billion K-line records from 45 global exchanges, enabling it to learn nuanced temporal and cross-asset representations. Kronos excels in a zero-shot setting across a diverse set of financial tasks. On benchmark datasets, Kronos boosts price series forecasting RankIC by 93% over the leading TSFM and 87% over the best non-pre-trained baseline. It also achieves a 9% lower MAE in volatility forecasting and a 22% improvement in generative fidelity for synthetic K-line sequences. These results establish Kronos as a robust, versatile foundation model for end-to-end financial time series analysis. Our pre-trained model is publicly available at https://github.com/shiyu-coder/Kronos

## I. Introduction

A. **Application of TSFM into Financial:** Within this expanding research landscape, financial markets stand out as a critical and challenging application area for TSFMs, given their inherent data richness, highfrequency observations, and complex, non-stationary temporal dynamics. At the core of this domain are K-line sequences, multivariate time series derived from candlestick charts that record **Open, High, Low,** and **Close**
**prices,** along with trading **Volume** and **Amount** (Turnover)
over fixed intervals **(OHLCVA)**.

B. **Những vấn đề của việc áp dụng TSFMs vào financial K-line data**:

- K-line sequences exhibit unique statistical properties—such as **low signal-to-noise ratios, strong non-stationarities, and intricate, high-order dependencies among OHLCVA attributes** (Zhang and Hua 2025; Baidya and Lee 2024)—that are **often misaligned with the inductive biases of generic TSFMs**

- Second, the financial domain has largely been underserved by mainstream TSFM research; **financial sequences constitute a minor fraction of pre-training corpora for most existing TSFMs** (Das et al. 2024; Gao et al. 2024; Xiaoming et al. 2025) , and the spectrum of downstream tasks critical to quantitative finance—**spanning volatility estimation, synthetic sequence generation, and risk management**—remains largely unaddressed

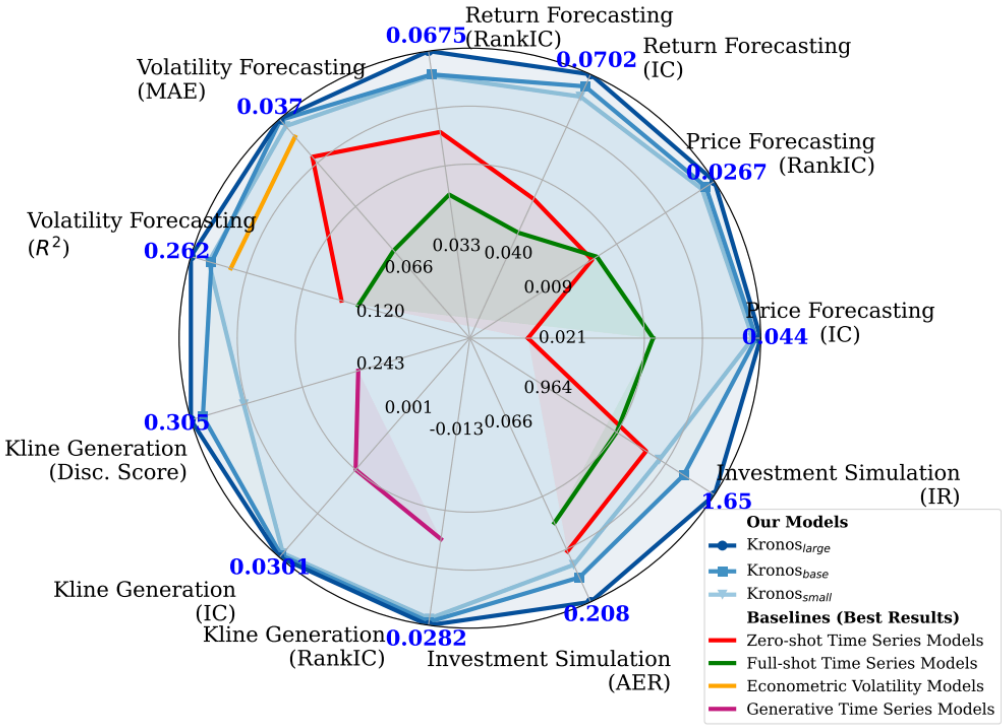
## C. Solution

**Kronos**, a unified, scalable pre-training framework designed specifically for financial K-line data:

- employs a specialized tokenizer to discretize K-line inputs into a sequence of compact tokens
- undergoes autoregressive pre-training on over **12 billion K-line records** drawn from over **45 global markets** and **7 temporal granularities.**

**Efficacy's Validations:**

- Price series forecasting: Kronos đạt state-of-the-art, tăng RankIC **93% so với TSFM tốt nhất** và **87% so với baseline không pretrain tốt nhất**
- Volatility forecasting: **MAE thấp hơn 9%**
- Synthetic K-line generation: **generative fidelity cải thiện 22%**
- Tổng quan: kết quả cho thấy hiệu quả rộng và tính đa dụng của Kronos trên nhiều tác vụ định lượng tài chính



# II. Preliminary

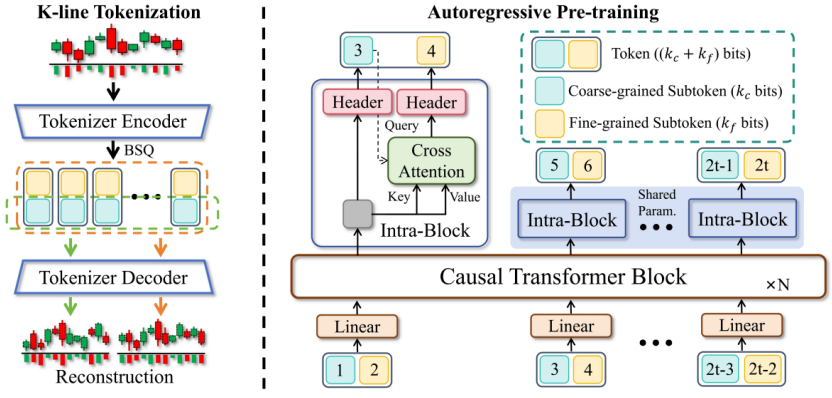
- Mỗi quan sát K-line tại thời điểm t được biểu diễn thành vector **xt** ∈ **RD**; trong bài này **D=6** tương ứng **OHLCVA**.
- Cho chuỗi lịch sử **x(1:T) = (x1, x2, . . . , xT )**, mục tiêu là dự báo **H** bước tiếp theo **xˆT +1:T +H = (xˆT +1, xˆT +2, . . . , xˆT +H)** theo bài toán forecasting chuẩn.
- Thay vì dự báo trực tiếp trên số thực, Kronos lượng tử hóa mỗi **xt** thành token rời rạc **bt** bằng một codebook học được C, biến chuỗi liên tục **x1:T = (x1, . . . , xT )** thành chuỗi token **b1:T = (b1, . . . , bT ).**
- Bài toán dự báo được chuyển thành mô hình hóa chuỗi token theo cơ chế tự hồi quy:

$$p(\mathbf{b}_{T+1:T+H} \mid \mathbf{b}_{1:T}) = \prod_{h=1}^{H} p\big(b_{T+h} \mid \mathbf{b}_{1:T+h-1}\big).$$

- Cách biểu diễn rời rạc này giúp mở rộng tự nhiên sang các tác vụ mang tính sinh: dự báo giá, suy ra volatility từ quỹ đạo dự báo, và sinh chuỗi K-line tổng hợp bằng cách generate token rồi decode về OHLCVA.

# III. Methodology

**A.** Kronos abstracts financial K-line sequences as a discrete language and implements this via a two-phase framework:
**(1) K-line Tokenization:** A specialized Transformer tokenizer quantizes each OHLCVA bar into a discrete token via a learnable codebook, structured as coarse and fine subtokens enforced by hierarchical reconstruction to capture multi-scale information.
**(2) Autoregressive Pre-training:** A decoder-only Transformer is trained on the token sequences with next-token prediction, sequentially generating coarse then fine subtokens conditioned on history to learn a high-fidelity hierarchical representation of market dynamics.



## B. K-line Tokenization

### Mục tiêu

- Chuyển chuỗi K-line liên tục **x = (x1, . . . , xT ), xt ∈ RD** (OHLCVA) thành chuỗi token rời rạc **b=(b1,…,bT)** để phục vụ mô hình hóa tự hồi quy ở Phase 2.
- Xây dựng token có cấu trúc phân cấp coarse-to-fine nhằm biểu diễn động lực thị trường đa thang đo.

### Kiến trúc Tokenizer

- Transformer-based autoencoder gồm 3 khối chính:
  - Encoder $E_{enc}$ mã hóa mỗi K-line item thành latent liên tục $\xi t$.
  - Quantizer Q: lượng tử hóa $\xi t$ thành mã nhị phân theo BSQ.
  - Decoder $E_{dec}$: giải mã token để tái tạo lại tín hiệu gốc và tạo tín hiệu học cho tokenizer.

### Quantization bằng Binary Spherical Quantization (BSQ)

- BSQ ánh xạ latent liên tục $\xi t$ thành mã nhị phân k-bit:
  - $bt \in \{-1, 1\}^k$ thông qua chiếu lên các hyperplane học được.
- Động cơ lựa chọn:
  - Biểu diễn rời rạc giúp bài toán downstream chuyển thành dự đoán phân loại/token thay vì hồi quy trực tiếp.
  - k lớn tăng sức biểu đạt nhưng kéo theo vocabulary kích thước $2^k$, gây nặng cho mô hình tự hồi quy.

### Factorization token thành subtokens (n = 2)

- Để tránh vocabulary khổng lồ $2^k$, tách k-bit code thành 2 không gian con:
  - Coarse subtoken $b_t^c \in \{-1, 1\}^{k/2}$
  - Fine subtoken $b_t^f \in \{-1, 1\}^{k/2}$
  - Token đầy đủ $b_t = [b_t^c, b_t^f]$
- Hệ quả tính toán:

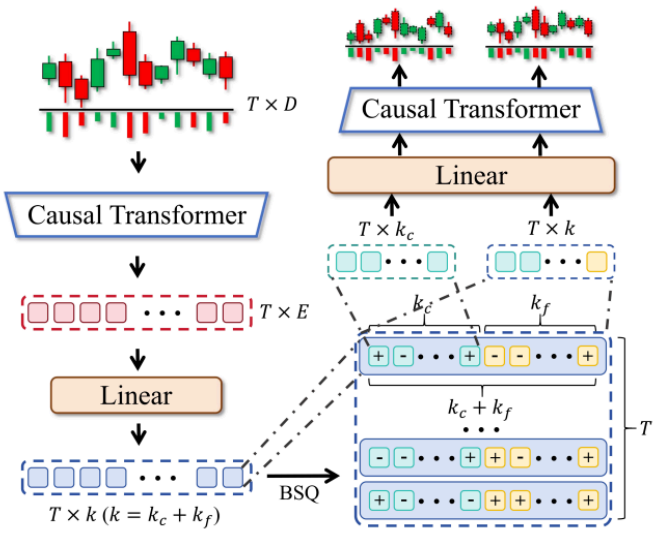
- Thay 1 dự đoán trên $2^k$ bằng 2 dự đoán tuần tự trên $2^{k/2}$, giảm mạnh chi phí tham số và tính toán của embedding/projection trong Phase 2.

**Hàm mất mát phân cấp để ép cấu trúc coarse-to-fine**

- **Mục tiêu tổng:**
  - Ltokenizer = Lcoarse + Lfine + λLquant
- **Thành phần:**
  - Lcoarse = $E\|x - E_{dec}(b^c)\|^2$
    - Buộc coarse subtoken học phần cấu trúc chính, tái tạo mức thô.
  - Lfine = $E\|x - E_{dec}(b)\|^2$
    - Buộc token đầy đủ (coarse+fine) tái tạo chất lượng cao; fine subtoken học phần residual để tinh chỉnh.
  - Lquant: loss của BSQ, regularize khoảng cách giữa latent ξ và binary code b để ổn định lượng tử hóa.

**Ý nghĩa thiết kế**

- Token thu được mang tính phân cấp:
  - Coarse subtoken đại diện thông tin thô, đóng vai trò scaffold.
  - Fine subtoken bổ sung chi tiết, mã hóa phần bù để nâng chất lượng biểu diễn.
- Tạo điều kiện cho Phase 2 dự đoán theo thứ tự coarse rồi fine, mô hình hóa phụ thuộc nội bộ token một cách có chủ đích.



## C. Hierarchical Autoregressive Modeling

### Mục tiêu và phân rã xác suất

- Đầu vào là chuỗi token rời rạc **b={b1,…,bT}** sau tokenization, với mỗi token có cấu trúc phân cấp
    - $b_t = [b_t^c, b_t^f]$ gồm coarse subtoken và fine subtoken
- Dùng decoder-only Transformer $E_{ar}$ với causal attention để mô hình hóa phân phối chung của chuỗi:
    - $p(b) = \prod_{t=1}^{T} p(b_t \mid b_{<t})$, trong đó $b_{<t}$ là toàn bộ token trước thời điểm t
- Áp dụng chain rule để hiện thực hóa phụ thuộc coarse-to-fine ngay trong xác suất điều kiện:
    - $p(b_t \mid b_{<t}) = p(b_t^c \mid b_{<t}) \cdot p(b_t^f \mid b_{<t}, b_t^c)$
- Ý nghĩa thiết kế:
    - Coarse được sinh trước để đóng vai trò scaffold
    - Fine được sinh sau để mã hóa phần residual, hoàn thiện token ở mức chi tiết

### Xây dựng vector đầu vào theo từng time step

- Với mỗi thời điểm i, hai subtokens được embedding độc lập bằng hai bảng embedding khác nhau:
    - $e_c(b_i^c)$ cho coarse, $e_f(b_i^f)$ cho fine
- Ghép hai embedding và chiếu tuyến tính để tạo vector đầu vào hợp nhất:
    - $v_i = W_{fuse}([e_c(b_i^c); e_f(b_i^f)])$
    - [ ; ] là phép nối vector; $W_{fuse}$ là ma trận học được để đưa về latent space của mô hình
- Chuỗi đầu vào {v1,…,vt−1} được đưa qua Ear(causal) để thu hidden state theo ngữ cảnh; hidden đại diện cho lịch sử đến t−1được ký hiệu:
    - $h_t = E_{ar}(v_{<t})$ (paper mô tả là final hidden state khi xử lý $b_{<t}$)

### Dự đoán coarse subtoken

- Từ history vector $h_t$, dùng head tuyến tính $W_c$ tạo logits và softmax để ra phân phối coarse:
    - $p(b_t^c \mid b_{<t})$ = softmax($W_c h_t$)

- Đây là bước dự báo mức thô ở thời điểm t, độc lập với fine.

**Dự đoán fine subtoken có điều kiện trên coarse**

- Để mô hình hóa đúng điều kiện $p(b_t^f \mid b_{<t}, b_t^c)$, paper cập nhật ngữ cảnh bằng coarse đã dự đoán:
  - Trong training, coarse không teacher-forcing; dùng $\hat{b}_t^c$ được sample từ $p(b_t^c \mid b_{<t})$
  - Mục đích: giảm exposure bias, khớp với inference multi-step (không có ground-truth token tương lai)
- Cơ chế cập nhật dùng cross-attention:
  - Query: embedding của coarse dự đoán $e_c(\hat{b}_t^c)$
  - Key/Value: history vector $h_t$
  - $h_t^{update}$ = CrossAttn( q = $e_c(\hat{b}_t^c)$, k = v = $h_t$)
- Từ $h_t^{update}$, dùng head $W_f$ để dự đoán phân phối fine:
  - $p(b_t^f \mid b_{<t}, b_t^c)$ = softmax($W_f h_t^{update}$)

**Hàm mục tiêu huấn luyện (negative log-likelihood)**

- Tối ưu log-likelihood của cả hai bước dự đoán tại mọi thời điểm:

$$\mathcal{L}_{\text{ar}} = -\mathbb{E}_{\mathbf{b}\sim\mathcal{D}} \sum_{t=1}^{T} \Big[\log p(b_t^c|\mathbf{b}_{<t}) + \log p(b_t^f|\mathbf{b}_{<t}, b_t^c)\Big]$$

- $D$ là phân phối dữ liệu tokenized (từ corpus K-line sau Phase 1).

**Quy trình inference**

Tại mỗi time step **t**:

1. Từ lịch sử token đã có $b_t$, Transformer cho ra $h_t$
2. Sinh coarse: sample hoặc argmax từ $p(b_t^c \mid b_{<t})$
3. Cập nhật ngữ cảnh bằng cross-attn với coarse vừa sinh

4. Sinh fine từ $p(b_t^f \mid b_{<t}, b_t^c)$
5. Ghép lại thành $b_t = [b_t^c, b_t^f]$ , append vào chuỗi và lặp sang t+1

## D. Model Pre-training

### Dataset

- Xây dựng bộ dữ liệu K-line tài chính quy mô lớn và chất lượng cao phục vụ tiền huấn luyện.
- Bối cảnh: dữ liệu tài chính toàn diện và được làm sạch kỹ lưỡng còn hạn chế so với các bộ dữ liệu chuỗi thời gian tổng quát.
- Quy mô dữ liệu:
  - Hơn 12 tỷ quan sát.
  - 7 tần suất lấy mẫu.
  - Nhiều loại tài sản.
  - 45 sàn giao dịch toàn cầu.
- Thiết kế pipeline làm sạch dữ liệu chuyên biệt cho K-line:
  - Phát hiện và loại bỏ các đoạn giá biến động bất thường.
  - Lọc các giai đoạn thanh khoản thấp hoặc không có giao dịch kéo dài.
  - Đảm bảo tính nhất quán và độ tin cậy của chuỗi OHLCVA.

### Model Training

- Thiết kế dựa trên nguyên lý scaling laws quan sát trong huấn luyện LLM.
- Huấn luyện ba biến thể Kronos với số lượng tham số tăng dần.
- Mô hình lớn nhất gần 0.5 tỷ tham số.
- Mục tiêu: cân bằng giữa hiệu năng dự báo và chi phí suy luận.
- Giới hạn độ dài ngữ cảnh tối đa ở 512 token do ràng buộc tài nguyên và yêu cầu triển khai thực tế.
- Hỗ trợ linh hoạt các chân trời dự báo thông qua dữ liệu đa tần suất:
  - Dữ liệu 1 phút cho dự báo ngắn hạn.
  - Dữ liệu ngày cho dự báo trung và dài hạn như tuần hoặc tháng.

| | Layers | $\mathbf{d}_{\text{model}}$ | $\mathbf{d}_{\text{ff}}$ | Heads | Vocab. ($2^k$) | Params |
|---|---|---|---|---|---|---|
| $\text{Kronos}_{small}$ | 8 | 512 | 1024 | 8 | 20 | 24.7M |
| $\text{Kronos}_{base}$ | 12 | 832 | 2048 | 16 | 20 | 102.3M |
| $\text{Kronos}_{large}$ | 18 | 1664 | 3072 | 32 | 20 | 499.2M |

**Inference**

- Sinh chuỗi token tương lai theo cơ chế tự hồi quy, tương tự mô hình sinh văn bản.
- Điều khiển tính ngẫu nhiên bằng:
  - Temperature scaling.
  - Top-p sampling.
- Xác suất lấy mẫu token i từ logits z:
  - $p_i \propto exp(z_i/T)$, với T là tham số temperature.
- Đối với các tác vụ yêu cầu độ chính xác cao:
  - Sinh nhiều quỹ đạo dự báo tương lai bằng Monte Carlo rollouts.
  - Giải mã về giá trị liên tục và lấy trung bình các quỹ đạo.
  - Tăng độ ổn định và cải thiện chất lượng dự báo.
- Thực nghiệm cho thấy phương pháp này cải thiện nhất quán hiệu năng dự báo.

# IV. Experimentals

Mục tiêu của phần thực nghiệm là đánh giá toàn diện Kronos như một foundation model cho dữ liệu K-line tài chính, thông qua một bộ thí nghiệm bao phủ cả bài toán dự báo và sinh dữ liệu, đồng thời kiểm chứng khả năng chuyển hóa thành hiệu quả đầu tư thực tế.

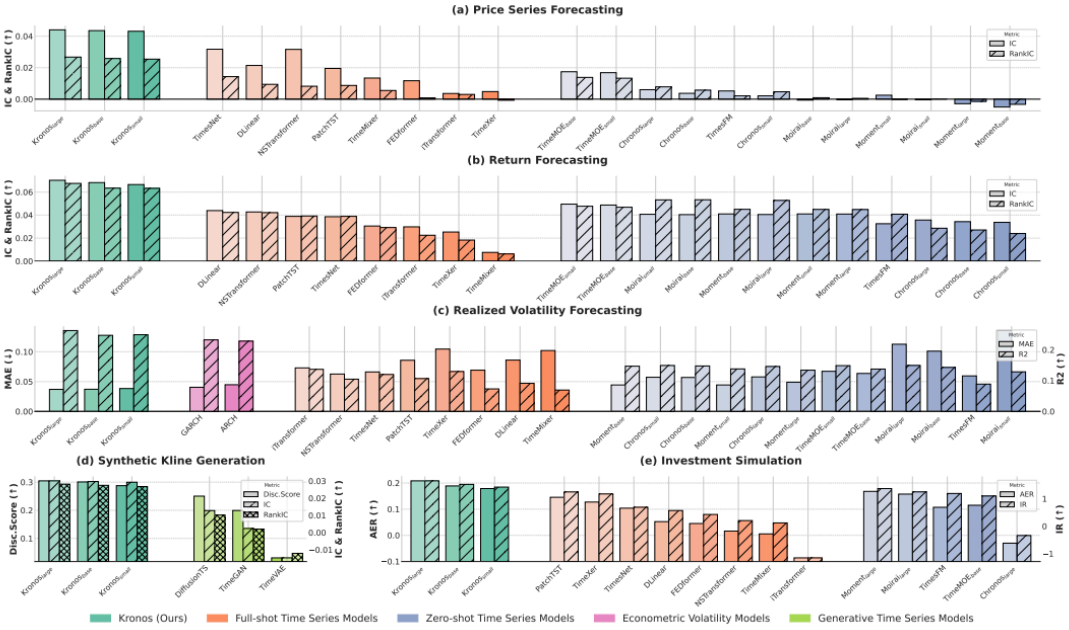



Figure 4: Main experimental results across five representative financial tasks. Subfigures (a-c) show forecasting performance on price series, returns, and realized volatility. Subfigure (d) displays generative model performance in terms of fidelity and usefulness. Subfigure (e) presents the investment simulation backtesting results.

## A. Experimental Setup

### Nhóm tác vụ đánh giá

Bộ thí nghiệm gồm năm tác vụ đại diện, chia thành ba nhóm chính:

**Tác vụ dự báo (Predictive tasks)**

- Price series forecasting
- Return forecasting
- Realized volatility forecasting

Các tác vụ này đo lường khả năng mô hình hóa động lực thị trường và chất lượng tín hiệu dự báo.

**Tác vụ sinh dữ liệu (Generative task)**

- Synthetic K-line generation

Đánh giá khả năng học phân phối dữ liệu và tái tạo động lực thị trường.

**Tác vụ mô phỏng đầu tư (Investment simulation)**

- Mô phỏng chiến lược long-only trên thị trường thực tế

Đánh giá khả năng chuyển hóa tín hiệu dự báo thành lợi nhuận đầu tư đo lường được.

## B. Hệ thống baseline

Kronos được so sánh với 25 mô hình baseline đại diện cho bốn hướng tiếp cận khác nhau:

**Full-shot time series models**
Các mô hình hiện đại huấn luyện trực tiếp trên downstream task, không tiền huấn luyện nền tảng.

**Zero-shot time series foundation models**
Các TSFM tiền huấn luyện đa miền, đánh giá theo thiết lập zero-shot trên dữ liệu tài chính.

**Econometric volatility models**
Các mô hình kinh tế lượng cổ điển như ARCH và GARCH cho bài toán volatility forecasting.

**Generative time series models**
Các mô hình sinh chuỗi như Diffusion-based, VAE-based và GAN-based.

Cách lựa chọn baseline bảo đảm so sánh công bằng giữa các paradigm: hồi quy liên tục, mô hình xác suất, mô hình foundation tổng quát và mô hình sinh chuyên biệt.

## C. Main Results

### Prediction Tasks

Kronos đạt state-of-the-art nhất quán trên cả ba tác vụ forecasting.

**Price series forecasting**

- RankIC cải thiện 93% so với TSFM mạnh nhất.
- Cải thiện 87% so với mô hình không tiền huấn luyện tốt nhất.

**Return forecasting**

- IC và RankIC đều vượt trội so với toàn bộ baseline.

**Realized volatility forecasting**

- Giảm MAE.
- Tăng $R^2$ so với các phương pháp cạnh tranh.

Một kết quả quan trọng là khi tăng kích thước mô hình từ small → base → large, hiệu năng cải thiện đều đặn. Điều này xác nhận tính hợp lệ của scaling laws trong bối cảnh foundation model cho chuỗi thời gian tài chính.

## D. Generative Tasks

Chất lượng dữ liệu sinh được đánh giá theo ba tiêu chí:

**Diversity**
Đánh giá mức độ bao phủ phân phối dữ liệu thật.

- Sử dụng t-SNE để chiếu dữ liệu thật và dữ liệu sinh về không gian 2D.

- Sử dụng Kernel Density Estimation để so sánh phân phối.
  Kronos cho mức chồng lấp phân phối cao hơn và bao phủ tốt hơn không gian dữ liệu thật.

**Fidelity**

Đánh giá tính chân thực của dữ liệu sinh.

- Sử dụng discriminative score, đo khả năng phân biệt giữa dữ liệu thật và dữ liệu sinh.
  Kronos đạt fidelity cao nhất, tức dữ liệu sinh khó bị phân biệt.

**Usefulness**

Đánh giá tính hữu dụng của dữ liệu sinh cho downstream task.

- Áp dụng giao thức Train-on-Synthetic, Test-on-Real.
- Huấn luyện mô hình dự báo trên dữ liệu sinh.
- Đánh giá IC và RankIC trên tập test thật.
  Kronos đạt kết quả cao nhất, chứng minh dữ liệu sinh không chỉ giống thật mà còn hữu ích cho dự báo.

Hiệu năng sinh dữ liệu cũng tăng khi kích thước mô hình tăng.

Table 2: Ablation study dissecting the architectural choices of Kronos. We compare our model against variants targeting different **Prediction Spaces** (continuous vs. discrete) with corresponding **Training Objectives**. *Direct-AR* serves as a standard regression baseline. *Prob-AR* evaluates the benefit of probabilistic modeling in the continuous space. *Kronos-Parallel* ablates our sequential subtoken design by predicting subtokens concurrently. Best results are in **bold**.

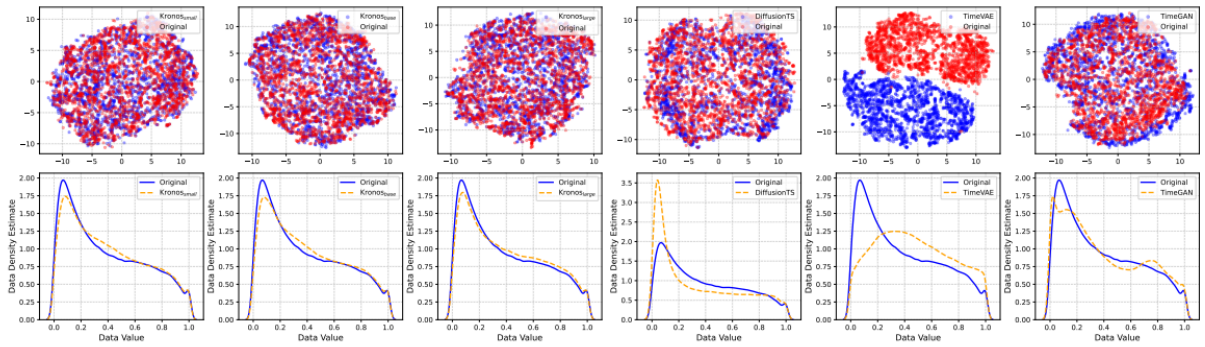



Figure 5: Visual comparison of generative models on the dataset of Shanghai Stock Exchange, 15-minute frequency. **Top row:** t-SNE embeddings of original (red) versus synthetic (blue) data. **Bottom row:** Kernel Density Estimates (KDE) of original versus synthetic data.

## E. Investment Simulation

Để kiểm chứng tính thực tiễn, tác giả thực hiện mô phỏng chiến lược long-only trên thị trường Chinese A-shares:

- Xếp hạng cổ phiếu theo tín hiệu dự báo của từng mô hình.
- Chọn top-k cổ phiếu để xây dựng danh mục.
- Tính toán hiệu suất theo thời gian.

Kronos đạt:

- Annualized Excess Return cao nhất.
- Information Ratio cao nhất.

Điều này cho thấy cải thiện về IC và RankIC thực sự chuyển hóa thành lợi nhuận đầu tư.

## F. Ablation Study

Ablation tập trung vào hai câu hỏi cốt lõi:

- Hiệu quả của framework rời rạc và tự hồi quy so với mô hình hóa liên tục.
- Ảnh hưởng của kích thước vocabulary.

### Analysis of Modeling Paradigms

So sánh bốn biến thể:

**Direct-AR**

- Không gian liên tục.
- Hồi quy trực tiếp với MSE.

**Prob-AR**

- Không gian liên tục.
- Mô hình xác suất với phân phối Student-t mixture để xử lý heavy-tail.

**Kronos-Parallel**

- Không gian rời rạc.
- Dự đoán coarse và fine subtokens đồng thời.

**Kronos (Sequential)**

- Không gian rời rạc.
- Dự đoán coarse trước, sau đó fine có điều kiện trên coarse.

Kết quả:

- Các mô hình rời rạc vượt trội so với mô hình liên tục.
- Dự đoán tuần tự coarse → fine tốt hơn dự đoán song song.

Kết luận:
Việc mô hình hóa phụ thuộc nội bộ giữa các subtokens là yếu tố quan trọng, và không gian rời rạc phù hợp hơn với đặc tính phân phối nặng đuôi và phi tuyến của dữ liệu tài chính.

| Model | Prediction Space | Training Objective | Price Series Forecasting | | Return Forecasting | | Volatility Forecasting | |
|---|---|---|---|---|---|---|---|---|
| | | | IC (↑) | RankIC (↑) | IC (↑) | RankIC (↑) | MAE (↓) | $R^2$ (↑) |
| Direct-AR | Continuous | Mean Squared Error (MSE) | 0.0212 | 0.0149 | 0.0416 | 0.0399 | 0.0565 | 0.1608 |
| Prob-AR | Continuous | Negative Log-Likelihood (NLL) | 0.0179 | 0.0102 | 0.0356 | 0.0329 | 0.0464 | 0.1383 |
| Kronos-Parallel | Discrete | Cross-Entropy | 0.0345 | 0.0226 | 0.0529 | 0.0505 | 0.0461 | 0.1784 |
| **Kronos**$_{small}$ | **Discrete** | **Cross-Entropy** | **0.0431** | **0.0254** | **0.0665** | **0.0622** | **0.0384** | **0.2490** |

Table 2: Ablation study dissecting the architectural choices of Kronos. We compare our model against variants targeting different **Prediction Spaces** (continuous vs. discrete) with corresponding **Training Objectives**. *Direct-AR* serves as a standard regression baseline. *Prob-AR* evaluates the benefit of probabilistic modeling in the continuous space. *Kronos-Parallel* ablates our sequential subtoken design by predicting subtokens concurrently. Best results are in **bold**.

### Impact of Vocabulary Size

Tác giả phân tích ảnh hưởng của kích thước vocabulary 2^k:

- Vocabulary lớn hơn → biểu diễn mịn hơn.
- Giảm quantization error.
- Cải thiện reconstruction quality.
- Đồng thời cải thiện IC, RankIC và R² trong forecasting.

Kết quả cho thấy độ chính xác biểu diễn ở phase tokenization có ảnh hưởng trực tiếp đến hiệu năng downstream.

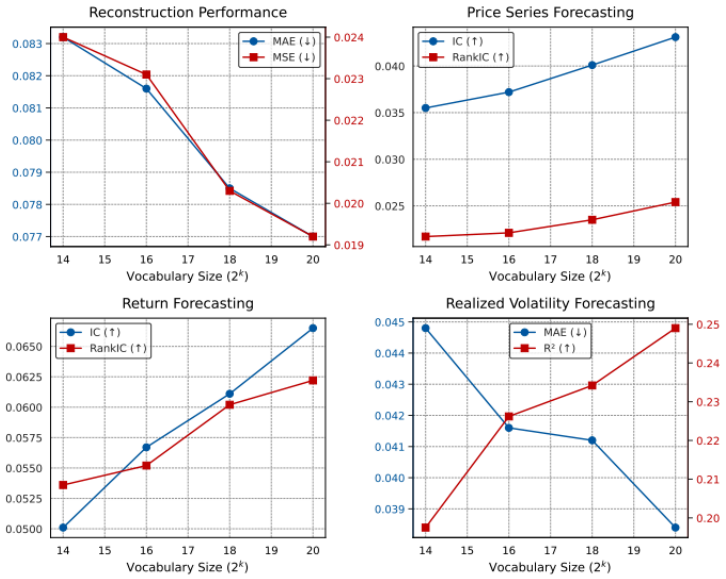



Figure 6: Impact of vocabulary size on model performance. We plot reconstruction quality and downstream forecasting performance as vocabulary size increases.

## G. Test-Time Scaling

Một ưu điểm nổi bật của framework sinh xác suất là khả năng cải thiện dự báo tại inference mà không cần huấn luyện lại.

Cơ chế:

- Từ cùng một context, mô hình sinh nhiều quỹ đạo tương lai khác nhau.
- Trung bình các quỹ đạo để tạo dự báo cuối cùng.

Quan sát thực nghiệm:

- IC và RankIC tăng khi tăng số lượng sample.
- Averaging giảm phương sai do stochastic sampling.
- Dự báo ổn định và robust hơn.

Hàm ý thực tiễn:

- Có thể đánh đổi giữa chi phí tính toán tại inference và độ chính xác mong muốn.
- Không cần thay đổi tham số mô hình để nâng cao hiệu năng.

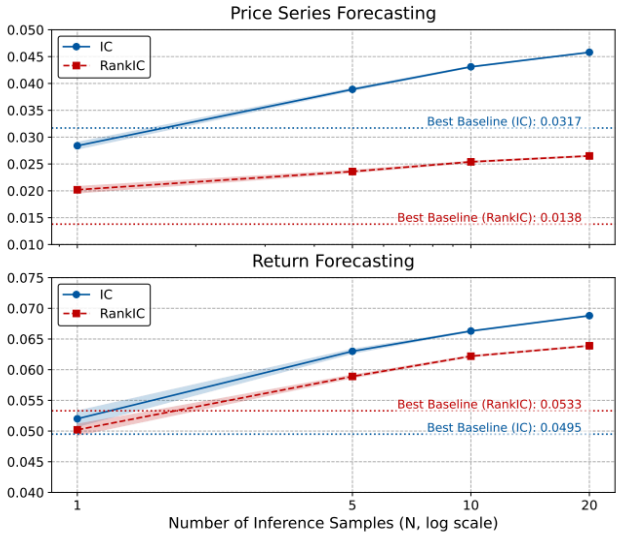



Figure 7: Impact of the number of inference samples (N) on forecasting performance.The lines represent the mean performance over 5 runs with different random seeds, while the shaded areas indicate the standard deviation.

**Kết luận từ Section 4**

Phần thực nghiệm chứng minh:

- Kronos vượt trội trên cả dự báo và sinh dữ liệu.
- Discrete hierarchical autoregressive modeling phù hợp với tài chính hơn hồi quy liên tục.
- Scaling law vẫn giữ vai trò quan trọng trong foundation model cho chuỗi thời gian.
- Biểu diễn rời rạc chất lượng cao dẫn đến cải thiện downstream rõ rệt.
- Mô hình có khả năng chuyển hóa tín hiệu dự báo thành hiệu quả đầu tư thực tế.
- Test-time ensembling cung cấp một cơ chế mở rộng hiệu năng linh hoạt.

# V. Conclusion

In this work, we introduce Kronos, a foundation model specifically designed for financial K-line sequences. Kronos employs a novel two-stage framework, where an instancebased tokenizer first discretizes continuous market data into hierarchical coarse-to-fine tokens, which are then modeled by a large autoregressive Transformer. Comprehensive empirical evaluations demonstrate that Kronos establishes new state-of-the-art benchmarks in price series forecasting, as well as in other relevant applications such as synthetic Kline generation and volatility forecasting, significantly outperforming existing TSFMs and other baselines. These results position Kronos as a robust and versatile foundation for a range of applications in quantitative finance.