AI4Trade — Feature Engineering & Design Rationale (31 Features)

Repository Companion Document — Indian Westerlies

# 1) Objective

This document enumerates the 31 deterministic features we engineered for monthly bilateral trade forecasting at the HS‑6 level, with design rationale grounded in trade economics and practical forecasting needs. All features are computed per series (origin, destination, hs6, trade\_flow, month) using only information available up to t−1 to avoid leakage. Predictions are later aggregated to HS‑4 for submission.

# 2) Design Principles

• Deterministic, lag‑safe: Every feature is a function of historical data ≤ t−1.  
• Segment‑aware horizons: CHN uses h=2; USA uses h=3.  
• Interpretability: Each feature ties to an economic intuition (trend, seasonality, volatility, macro/supply‑chain).  
• Parsimony: 31 features selected via ablation and validation for strong sMAPE performance.

# 3) Feature Families (with Rationale)

## A) Temporal & Calendar (3)

1. month\_num — Month number (1–12) to encode seasonality (holidays, CNY, shopping seasons).

2. quarter — Calendar quarter (1–4) to capture quarterly production/ordering cycles.

3. month\_id — Monotone index for long‑run drift/trend capturing structural changes.

## B) Lagged Levels (6)

4. lag1 — Previous month level; contracts/inertia in shipments.

5. lag2 — Two‑month memory; aligns with h=2 for CHN.

6. lag3 — Three‑month persistence; aligns with h=3 for USA.

7. lag6 — Half‑year memory; supplier re‑ordering cadence.

8. lag10 (CHN) / lag9 (USA) — Horizon‑tuned mid‑range memory.

9. lag12 — Annual seasonality anchor; year‑over‑year echo.

## C) Moving Averages (3)

10. ma3 — Short‑term smoothing; reduces month‑to‑month noise.

11. ma6 — Medium‑term smoothing; inventory smoothing effects.

12. ma12 — Yearly smoothing; baseline for weighting/normalization elsewhere.

## D) Volatility & Momentum (3)

13. roll\_std6 — Six‑month rolling standard deviation; dispersion/uncertainty proxy.

14. pctchg1 — Month‑over‑month % change; near‑term momentum.

15. pctchg3 — Three‑month % change; medium‑horizon trend acceleration.

## E) Cross‑Flow Dynamics (3)

16. cross\_flow\_lag1 — Opposite flow level at t−1 for same origin–destination–HS (exports↔imports), computed at HS‑4 to increase signal density.

17. cross\_flow\_ma3 — Three‑month average of the opposite flow; smoother co‑movement.

18. cross\_flow\_lag13 — Year‑lagged opposite flow; slow substitution/complement effects.

## F) Macro & Supply‑Chain Pressure (4)

19. origin\_total\_exports\_ma3 — Country‑level export momentum (MA3), capturing broad demand/competitiveness shifts.

20. origin\_total\_imports\_ma3 — Country‑level import momentum (MA3), proxy for intermediate‑goods appetite.

21. gscpi\_ma2\_shift — Global Supply Chain Pressure Index, 2‑month moving average shifted to avoid leakage (shift=1 for China segments, shift=2 for USA segments). Encodes freight, delivery times, and global bottlenecks.

22. oecd\_cli\_ma2\_shift — OECD Composite Leading Indicator, 2‑month moving average shifted (shift=1 for China, shift=2 for USA). Captures cyclical turning points in aggregate activity relevant for trade.

## G) Activity / Zero‑Handling (2)

23. was\_trade\_lag1 — Indicator if there was any trade at t−1; helps extensive‑margin modeling.

24. consec\_zero\_run — Length of the current zero‑trade run; distinguishes new vs dormant lines.

## H) Additional Deterministic Context (7)

25. hs\_section\_prefix — HS section/chapter proxy from hs6; coarse product taxonomy signal.

26. partner\_tier\_rank — Rank of partner by trailing 12‑month flow within product; market depth proxy.

27. zero\_share\_flag — Flag if series share of HS‑section is ~0 recently; prunes false positives.

28. yoy\_delta — Difference between level at t and t−12; simple YOY momentum cue (uses lagged values only).

29. month\_sin — Seasonal encoding via sin(2π·month\_num/12).

30. month\_cos — Seasonal encoding via cos(2π·month\_num/12).

31. weight\_sqrt\_ma12 — Stabilizing weight = √(ma12 + 1) for robust training and loss shaping.

# 4) Economic Intuition (One‑Liners)

• Lags capture contract stickiness and production‑to‑shipment latency.  
• Moving averages/volatility encode inventory smoothing and order batching.  
• Cross‑flow features connect exports–imports through input–output ties and re‑export chains.  
• Macro & supply‑chain (GSCPI/CLI) inject global conditions and turning points without leaking future info.  
• Activity flags model entry/exit at the extensive margin, common in granular HS lines.

# 5) Implementation Notes (Leak‑safe)

• All features computed with data ≤ t−1; targets are y(t+h) with h=2 (CHN) and h=3 (USA).  
• GSCPI/CLI are aligned monthly, then 2‑month MA applied, then shifted: +1 for China segments, +2 for USA segments.  
• Cross‑flow computed at HS‑4 to improve density; merged back to HS‑6 keys.  
• Minimum history requirements enforced for lag12/ma12; rows failing requirements get NA and are safely handled during training.

# 6) Integration & Outputs

Features are joined to the four segment masters on (origin, destination, hs6, trade\_flow, month). Outputs include explicit train/test split Parquets (with “\_final” suffix) for each segment and horizon:  
• features\_{segment}\_train\_h{2|3}\_final.parquet  
• features\_{segment}\_test\_h{2|3}\_final.parquet

# Appendix — 31 Features at a Glance

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Family | Computation (summary) | Why it helps |
| month\_num | Temporal | Month integer 1–12 | Seasonality |
| quarter | Temporal | Quarter 1–4 | Quarterly cycles |
| month\_id | Temporal | Monotone month index | Trend/drift |
| lag1 | Lag | y(t-1) | Contract inertia |
| lag2 | Lag | y(t-2) | Near‑term memory (CHN h=2) |
| lag3 | Lag | y(t-3) | Near‑term memory (USA h=3) |
| lag6 | Lag | y(t-6) | Mid‑term memory |
| lag10/lag9 | Lag | y(t-10) CHN / y(t-9) USA | Horizon‑tuned memory |
| lag12 | Lag | y(t-12) | YOY seasonality |
| ma3 | MA | mean(y[t-3..t-1]) | Noise reduction |
| ma6 | MA | mean(y[t-6..t-1]) | Stability |
| ma12 | MA | mean(y[t-12..t-1]) | Baseline |
| roll\_std6 | Volatility | std(y[t-6..t-1]) | Uncertainty |
| pctchg1 | Momentum | (y(t-1)-y(t-2))/y(t-2) | MoM momentum |
| pctchg3 | Momentum | (y(t-1)-y(t-4))/y(t-4) | 3‑month momentum |
| cross\_flow\_lag1 | Cross‑flow | Opposite flow t−1 at HS‑4 | Complement/substitute |
| cross\_flow\_ma3 | Cross‑flow | MA3 of opposite flow | Smoother co‑movement |
| cross\_flow\_lag13 | Cross‑flow | Opposite flow t−13 | YOY coupling |
| origin\_total\_exports\_ma3 | Macro | MA3 of exports (origin) | Aggregate momentum |
| origin\_total\_imports\_ma3 | Macro | MA3 of imports (origin) | Intermediate‑goods demand |
| gscpi\_ma2\_shift | Supply‑chain | MA2 of GSCPI, shifted +1 (CHN) / +2 (USA) | Logistics pressure |
| oecd\_cli\_ma2\_shift | Macro‑leading | MA2 of CLI, shifted +1 (CHN) / +2 (USA) | Turning points |
| was\_trade\_lag1 | Activity | 1{y(t-1) > 0} | Extensive margin |
| consec\_zero\_run | Activity | Run length of zeros up to t−1 | Entry/exit dynamics |
| hs\_section\_prefix | Context | First 2–4 digits of HS | Taxonomy signal |
| partner\_tier\_rank | Context | Rank by trailing 12‑m flow | Market depth |
| zero\_share\_flag | Context | Near‑zero share recently | Suppress spurious fits |
| yoy\_delta | Context | y(t-1)-y(t-13) | YOY momentum |
| month\_sin | Seasonal enc. | sin(2π·month\_num/12) | Cyclic pattern |
| month\_cos | Seasonal enc. | cos(2π·month\_num/12) | Cyclic pattern |
| weight\_sqrt\_ma12 | Training aid | √(ma12+1) | Stabilize fitting |