SUPPLEMENTAL APPENDIX for ”A multimodal deep reinforcement learning framework for multi-period inventory decision-making under demand uncertainty”

Yu-Xin Tian[a], Chuan Zhang[a,*]

[a]*School of Business Administration, Northeastern University, Shenyang, 110169, China*

Appendix A. The time difference correlation analysis (TDCA) method

This section introduces the principles and process of feature selection using the TDCA method. Denote the total candidate feature number as N_f, let L ($L > 0$) represent the lag period of a feature sequence ahead of the arrival time, and $L_{\max}$ be the maximum lag period ($L \leq L_{\max}$). The expression of the candidate feature set is

$$\mathbb{F}_L = \left\{\mathbf{F}_{1,L}, \mathbf{F}_{2,L}, \ldots, \mathbf{F}_{n,L}, \ldots, \mathbf{F}_{N_f,L}\right\}, L = 1, 2, \ldots, L_{\max}, \tag{A.1}$$

where $\mathbf{F}_{n,L}$ represents a feature sequence.

We assess the correlation between the target and the feature sequences at various lag periods to optimally choose features for model training. Metrics for correlation measurement encompass the distance correlation coefficient (Székely et al., 2007), Pearson correlation coefficient (Rodgers and Nicewander, 1988), and Spearman correlation coefficient (Myers et al., 2010) and the copula entropy (Ma, 2021; Schnaubelt, 2022). We extensively employ these correlation coefficients, and the assessment criterion is determined by taking the maximum of their absolute values. Let $\mathbf{D}$ represent the demand sequence ending at the arrival time. The formula for the absolute correlation is expressed as

$$f_R\left(\mathbf{F}_{n,L}, \mathbf{D}\right) = \max_{f \in \{\text{Distance,Pearson,Spearman,CopEnt}\}} \left|f\left(\mathbf{F}_{n,L}, \mathbf{D}\right)\right|, \tag{A.2}$$

where ”Distance”, ”Spearman”, ”Pearson”, and ”CopEnt” is the function of the distance correlation, Pearson, Spearman and the copula entropy, respectively. The implementation process are as follows:

First, initiate the screening process for candidate features by computing the absolute correlation between each feature sequence with various lag periods before the arrival time and the demand sequence at the arrival time, utilizing Eq. (A.2). Establish an appropriate screening threshold $\hat{r}_1$, as the screening criterion, representing the minimum correlation between the demand and the selected features. Features meeting the condition $f_R\left(\mathbf{F}_{n,L}, \mathbf{D}\right) \geq \hat{r}_1$ are chosen, and the set of selected features is defined as $\mathcal{X}$.

Second, to address multicollinearity among the initially screened features, the Pearson correlation coefficient is employed. The removal criterion is defined by the threshold $\hat{r}_2$, which represents the maximum permissible correlation among the chosen features. Feature pairs with high linear correlation in set $\mathcal{X}$ are identified, and the feature with the relatively weaker correlation to the demand sequence is removed from the set. Negative correlations are also considered in this procedure, as highly anti-correlated input features are also collinear.

Finally, repeat the second step, until the Pearson correlation between any two features in set $\mathcal{X}$ is no longer greater than $\hat{r}_2$. Consequently, the features in set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_p}\}$ represent the selected appropriate features, with the feature count N_p.

We extract the values of each feature in the features set $\mathcal{X}$ at time t, represented as $x_{n,t} \in \mathbf{x}_n$, to form the demand-related feature vector at time t, expressed as

$$\mathbf{X}_t = \left(x_{1,t}, x_{2,t}, \ldots, x_{N_p,t}\right). \tag{A.3}$$

Here, the screening thresholds and relevant removal criteria are hyperparameters determined through repeated experiments.

Appendix B. Transformer

The Transformer model, introduced by Vaswani et al. (2017), is a machine learning model designed for natural language processing tasks, with the self-attention mechanism as its core component. The structure of the Transformer is shown in Fig. B.1.

Transformer adopts a Seq2Seq structure composed of an encoder and a decoder. The encoder maps the input sequence into a fixed-length vector representation, and the decoder transforms the fixed-length vector into an output sequence. Both the encoder and the decoder consist of multiple identical layers, each containing a self-attention mechanism and a feed-forward neural network. The self-attention mechanism assigns a weight to each position by computing the similarity between that position and all others in the input sequence, thereby computing a weighted average that incorporates information from the entire sequence. The principles of each component are described as follows:

(1) *Self-attention mechanism*

The self-attention mechanism is the core component of Transformer. It computes the similarity between each position (word) in the input sequence and all other positions to assign weights, thus generating a weighted average for each word vector that incorporates context. The specific computation process is as follows:

Query, key, and value matrices: The input sequence X is transformed by three different linear layers to generate the query matrix Q, the key matrix K, and the value matrix V:

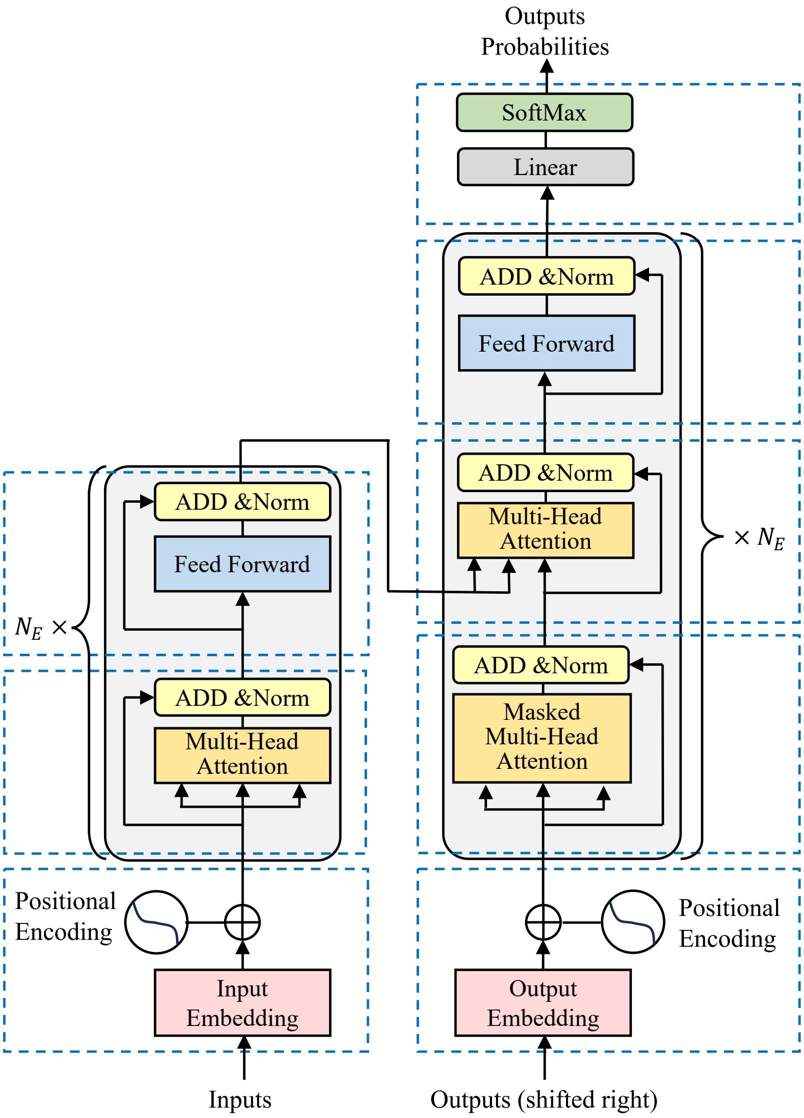

Fig. B.1. The structure of Transformer

$$Q = XW^Q, K = XW^K, V = XW^V, \tag{B.1}$$

where X represents the input sequence, and W^Q, W^K, and W^V are trainable weight matrices.

Attention scores: The dot product is computed between the query matrix Q and the key matrix K. Since the dimensionality d_K of the key vectors affects Transformer performance, the result is scaled by $\sqrt{d_K}$, and then normalized using the SoftMax function:

$$\text{Attention}(Q, K, V) = \text{soft max}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_K}}\right)V. \tag{B.2}$$

Multi-head attention mechanism: Multiple independent self-attention heads are introduced to enhance the model's capability, with each head operating in a different subspace. The formula for multi-head attention Mhead is as follows:

$$\mathrm{Mhead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_{N_h})W^O, \tag{B.3}$$

where $\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and W^O is a linear transformation matrix. The number of attention heads N_h is set to 4 in our experiment. Each head has its own set of weight matrices W_i^Q, W_i^K, and W_i^V.

(2) *Feed-forward neural network (FFN)*

The output at each position from the self-attention mechanism is processed by a feed-forward neural network, which consists of two linear transformations and a ReLU activation function $\max(0, \cdot)$:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2. \tag{B.4}$$

(3) *Residual connections and layer normalization*

To alleviate the vanishing gradient problem in deep networks, Transformer applies residual connections after both the self-attention and feed-forward sublayers, followed by layer normalization. The output of each sublayer is:

$$\mathrm{LayerNorm}(x + \mathrm{Sublayer}(x)), \tag{B.5}$$

where $\mathrm{Sublayer}(x)$ represents the function implemented by that sublayer.

(4) *Encoder and decoder*

Overall, Transformer consists of stacked encoder and decoder components. The encoder maps the input sequence into a fixed-length vector representation, and the decoder transforms this representation into the output sequence. Both the encoder and decoder are composed of N_E identical layers, each including multi-head self-attention and a feed-forward neural network. The computation steps of the encoder layer are:

$$x := \mathrm{LayerNorm}(x + \mathrm{FFN}(x)), \tag{B.6}$$

$$\mathrm{EncoderLayer}(x) = \mathrm{LayerNorm}(x + \mathrm{MHead}(x, x, x)). \tag{B.7}$$

In addition to the two sublayers in the encoder layer, the decoder layer includes a masked multi-head self-attention mechanism sublayer MaskedMHead. Similar to the encoder, residual connections and layer normalization are applied around each sublayer. The decoder layer computations are:

$$y := \mathrm{LayerNorm}(y + \mathrm{FFN}(y)), \tag{B.8}$$

$$y := \mathrm{LayerNorm}(y + \mathrm{MHead}(y, \varrho, \varrho)), \tag{B.9}$$

$$\mathrm{DecoderLayer}(y, \varrho) = \mathrm{LayerNorm}(y + \mathrm{MaskedMHead}(y, y, y)), \tag{B.10}$$

where ϱ denotes the output of the encoder layer.

(5) *Positional encoding*

Since Transformer lacks recurrence and convolution, it must be supplied with information about the positions of tokens in a sequence. To provide this, positional encodings are added to the input/output embeddings at the bottom of the encoder and decoder stacks. These encodings are defined using sine and cosine functions:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/v}\right), \tag{B.11}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/v}\right), \tag{B.12}$$

where *pos* represents the position index, i is the dimension index, and v is the embedding dimension.

Appendix C. Candidate macroeconomic indicators and Baidu indices

The candidate features before being screened using the TDCA method include 86 macroeconomic indicators and keywords from the Baidu search indices, as shown in Table C.1 and Table C.2, respectively.

Table C.1. List of candidate macroeconomic indicators

No.	Name	No.	Name
1	China: CPI: YoY	44	China: Financial Institutions: Foreign Currency Loans Balance
2	China: CPI: Transport and Communications: Transportation Facility: YoY	45	China: Financial Institutions: New Foreign Total Loans
3	China: CPI: Transport and Communications: Fuels for Transport Facility: YoY	46	China: Financial Institutions: Total Deposits Balance: RMB
4	China: CPI: Transport and Communications: Use and Maintenance of Transport Facility: YoY	47	China: Financial Institutions: New RMB Deposits
5	China: CPI: MoM	48	China: Financial Institutions: New RMB Deposits: Households
6	China: CPI: Transport and Communications: Transportation Facility: MoM	49	China: Demand Deposit Interest Rate
7	China: CPI: Transport and Communications: Fuels for Transport Facility: MoM	50	China: Time Deposit Rate: 3M

(Continued on next page)

(Continued from previous page)

No.	Name	No.	Name
8	China: CPI: Transport and Communications: Use and Maintenance of Transport Facility: MoM	51	China: Time Deposit Rate: 6M
9	China: RPI: MoM	52	China: Short-term Loan Interest Rate: 6M (Incl.)
10	China: RPI: Transportation and Communication Appliances: MoM	53	China: Short-term Loan Interest Rate: 6M-1Y (Inclusive)
11	China: RPI: Fuels: MoM	54	Chinaedium and Long-term Lending Rates: 1Y-3Y (Inclusive)
12	China: PPI: Total Industry Products: YoY	55	Chinaedium and Long-term Lending Rates: 3Y-5Y (Inclusive)
13	China: PPI: Total Industry Products: MoM	56	Chinaedium and Long-term Lending Rates: Above 5Y
14	China: PPI: Consumer Goods: Durable Consumer Goods: YoY	57	China: SH and SZ Stock Markets: Total Market Value (A and B Shares)
15	China: PPI: Extraction of Petroleum and Natural Gas: MoM	58	China: SH and SZ Stock Markets: Total Stock Turnover
16	China: PPI: Manufacture of Rubber and Plastic Products: MoM	59	SSE: Average P/E Ratio
17	China: PPI: Manufacture of Automobile: MoM	60	SSE Conglomerates Index
18	China: Purchasing Price Index of Raw Material,Fuel and Power: MoM	61	CSI 300 Index
19	Purchasing Price Index of Raw Material,Fuel and Power: Fuel and Power: YoY	62	SZSE Component Index
20	China: CGPI: YoY	63	SSE T-Bond Index: Closing
21	China: CGPI: MoM	64	SSE Corporate Bond Index: Closing
22	China: Export Unit Value Index: HS2: Total Index	65	Futures Settlement Price (Active Contract): Deformed Steel Bar
23	China: Export Price Index: HS2: Class 17: Vehicles, Aircraft, Vessels and Associated Transport Equipment	66	Futures Settlement Price (Active Contract): Natural Rubber
24	China: Import Unit Value Index: HS2: Total Index	67	Futures Settlement Price (Active Contract): Fuel Oil

(Continued on next page)

(Continued from previous page)

No.	Name	No.	Name
25	China: Import Price Index: HS2: Class 17: Vehicles, Aircraft, Vessels and Associated Transport Equipment	68	Futures Settlement Price (Active Contract): Stainless Steel
26	China: Market Price: Gasoline (92#): China VI:MONTHLY:LAST	69	Futures Settlement Price (Continuous): Fuel Oil
27	China: Market Price: Gasoline (95#): China VI:MONTHLY:LAST	70	Futures Settlement Price (Continuous): Natural Rubber
28	China: Value of Imports and Exports: CNY	71	Futures Settlement Price (Continuous): Deformed Steel Bar
29	China: Trade Balance: CNY	72	Futures Settlement Price (Continuous): Stainless Steel
30	China: M0	73	China: Macro-economic Climate Index: Coincident Index
31	China: M1	74	China: Macro-economic Climate Index: Leading Index
32	China: M2	75	China: Macro-economic Climate Index: Lagging Index
33	China: M0: YoY	76	China: Surveyed Urban Unemployment Rate
34	China: M1: YoY	77	Surveyed Urban Unemployment Rate in 31 Big Cities and Towns
35	China: M2: YoY	78	China: Surveyed Urban Unemployment Rate: YoY
36	China: Financial Institutions: Total Loans Balance: RMB	79	China: Manufacturing PMI
37	China: Financial Institutions: Total Loans Balance: RMB: YoY	80	China: Manufacturing PMI: Production
38	China: Financial Institutions: New RMB Loans	81	China: Manufacturing PMI: Large Enterprises
39	China: Financial Institutions: New RMB Loans: Households	82	China: Manufacturing PMI: Medium-sized Enterprises
40	China: Financial Institutions: New RMB Loans: Households: Short-term	83	China: Manufacturing PMI: Small Enterprises
41	China: Financial Institutions: New RMB Loans: Households: Mid & Long-term	84	China: Consumer Confidence Index

(Continued on next page)

(Continued from previous page)

No.	Name	No.	Name
42	China: Financial Institutions: Short-term Loas Balance: RMB	85	China: Consumer Satisfaction Index
43	China: Financial Institutions: Mid & Long-term Loans Balance: RMB	86	China: Consumer Expectation Index

Table C.2. Baidu index keyword sets

Topics or products	**Baidu Index keywords**
Automobile industry related	汽车保险 (Car Insurance), 汽车官网 (Car Official Website), 汽车摇号 (Car Lottery), 汽车新闻 (Car News), 汽车点评 (Car Review), 汽车购置税 (Car Purchase Tax), 汽车销量 (Car Sales), 汽车销量排行 (Car Sales Ranking), 汽车销量排行榜 (Car Sales Ranking List)
Lavida	一汽大众 (Faw-Volkswagen), 大众 (Volkswagen), 大众4s店 (Volkswagen 4S Dealership), 大众朗逸价格 (Volkswagen Lavida Price), 大众朗逸怎么样 (How Is the Volkswagen Lavida), 大众朗逸报价 (Volkswagen Lavida Quote), 大众朗逸报价及图片 (Volkswagen Lavida Quote and Pictures), 大众汽车官网 (Volkswagen Official Website), 朗逸 (Lavida), 朗逸价格 (Lavida Price), 朗逸多少钱 (How Much Is Lavida), 朗逸怎么样 (How Is the Lavida), 朗逸报价 (Lavida Quote), 朗逸最新报价 (Latest Quote of Lavida), 朗逸汽车 (Lavida Car), 朗逸油耗 (Lavida Fuel Consumption), 朗逸论坛 (Lavida Forum), 大众朗逸 (Volkswagen Lavida)
Emgrand	吉利帝豪 (Geely Emgrand), 吉利帝豪怎么样 (How Is Geely Emgrand), 吉利 (Geely), 吉利汽车 (Geely Automobile), 吉利4s店 (Geely 4S Dealership), 帝豪 (Emgrand), 帝豪汽车 (Emgrand Automobile), 吉利帝豪报价 (Geely Emgrand Price)
Haval H6	哈弗h6 (Haval H6), 哈弗h6报价 (Haval H6 Quote), 哈弗h6怎么样 (How About Haval H6), 哈弗h6运动版 (Haval H6 Sport Edition), 哈弗h6油耗 (Fuel Consumption of Haval H6), 哈弗h6新款 (New Version of Haval H6), 哈弗 (Haval), 哈弗汽车 (Haval Cars), 哈弗官网 (Official Website of Haval)
Camry	凯美瑞 (Camry), 凯美瑞油耗 (Fuel Consumption of Camry), 凯美瑞怎么样 (How about Camry?), 凯美瑞论坛 (Camry Forum), 凯美瑞2.0 (Camry 2.0), 凯美瑞多少钱 (How Much Is Camry), 凯美瑞报价 (Camry Quote), 凯美瑞汽车 (Camry Automobile), 丰田凯美瑞 (Toyota Camry), 丰田凯美瑞报价 (Toyota Camry Quote)

Appendix D. Selected numerical features and their correlation coefficients

Tables D.1–D.4 present the demand-related feature selection results for the four products. In our experiments, the threshold parameters are set as follows: for the Lavida experiment, $\hat{r}_1 = 0.35$ and $\hat{r}_2 = 0.7$; for the Emgrand experiment, $\hat{r}_1 = 0.35$ and $\hat{r}_2 = 0.7$; for the Haval H6 experiment, $\hat{r}_1 = 0.45$ and $\hat{r}_2 = 0.7$; and for the Camry experiment, $\hat{r}_1 = 0.4$ and $\hat{r}_2 = 0.7$.

Table D.1. Structured feature selection results for the Lavida experiment

Feature Name	**Lag order**	**Correlation**
China: CPI: Transport and Communications: Transportation Facility: MoM	3	CopEnt=-0.366
China: CPI: Transport and Communications: Transportation Facility: MoM	6	Spearman=-0.411
China: CPI: Transport and Communications: Transportation Facility: MoM	9	CopEnt=-0.400
China: CPI: Transport and Communications: Use and Maintenance of Transport Facility: MoM	11	Pearson=0.377
China: RPI: Transportation and Communication Appliances: MoM	5	Distance=0.351
China: RPI: Transportation and Communication Appliances: MoM	6	Distance=0.460
China: PPI: Manufacture of Rubber and Plastic Products: MoM	1	CopEnt=-0.368
China: PPI: Manufacture of Rubber and Plastic Products: MoM	2	CopEnt=-0.378
China: PPI: Manufacture of Rubber and Plastic Products: MoM	4	CopEnt=-0.360
China: PPI: Manufacture of Rubber and Plastic Products: MoM	7	CopEnt=-0.414
China: PPI: Manufacture of Automobile: MoM	2	CopEnt=0.509
China: PPI: Manufacture of Automobile: MoM	3	CopEnt=0.582
China: PPI: Manufacture of Automobile: MoM	4	CopEnt=0.566
China: PPI: Manufacture of Automobile: MoM	5	CopEnt=0.479
China: PPI: Manufacture of Automobile: MoM	6	CopEnt=0.420
China: PPI: Manufacture of Automobile: MoM	7	CopEnt=0.491
China: PPI: Manufacture of Automobile: MoM	8	CopEnt=0.668
China: PPI: Manufacture of Automobile: MoM	9	CopEnt=0.553
China: PPI: Manufacture of Automobile: MoM	10	CopEnt=0.716
China: PPI: Manufacture of Automobile: MoM	11	CopEnt=0.733
China: Trade Balance: CNY	8	Distance=0.361
China: M2: YoY	2	CopEnt=-0.385
China: Financial Institutions: New RMB Loans	1	Pearson=-0.415
China: Financial Institutions: New RMB Loans: Households	2	Distance=0.355

(Continued on next page)

(Continued from previous page)

Feature Name	Lag order	Correlation
China: Financial Institutions: New RMB Deposits	1	Distance=0.471
China: Financial Institutions: New RMB Deposits	6	CopEnt=-0.394
China: SH and SZ Stock Markets: Total Stock Turnover	1	Spearman=-0.366
Futures Settlement Price (Continuous): Fuel Oil	7	CopEnt=-0.402
Futures Settlement Price (Continuous): Natural Rubber	4	Spearman=-0.453
Futures Settlement Price (Continuous): Stainless Steel	12	CopEnt=0.773
China: Manufacturing PMI	7	CopEnt=-0.359
大众朗逸报价 (Volkswagen Lavida Quote)	2	CopEnt=-0.357
大众汽车官网 (Volkswagen Official Website)	12	CopEnt=-0.372
朗逸价格 (Lavida Price)	8	CopEnt=-0.357
朗逸油耗 (Lavida Fuel Consumption)	1	Spearman=0.362
汽车摇号 (Car Lottery)	10	CopEnt=-0.365
汽车购置税 (Car Purchase Tax)	10	CopEnt=-0.357
汽车销量 (Car Sales)	4	CopEnt=-0.352

Table D.2. Structured feature selection results for the Emgrand experiment

Feature Name	Lag order	Correlation
China: CPI: MoM	11	Pearson=0.424
China: CPI: Transport and Communications: Transportation Facility: MoM	6	Distance=0.440
China: CPI: Transport and Communications: Transportation Facility: MoM	8	CopEnt=-0.374
China: CPI: Transport and Communications: Transportation Facility: MoM	11	CopEnt=-0.359
China: CPI: Transport and Communications: Use and Maintenance of Transport Facility: MoM	10	Distance=0.350
China: CPI: Transport and Communications: Use and Maintenance of Transport Facility: MoM	12	CopEnt=-0.352
China: RPI: Transportation and Communication Appliances: MoM	5	Distance=0.417
China: RPI: Transportation and Communication Appliances: MoM	6	Distance=0.490

(Continued on next page)

(Continued from previous page)

Feature Name	Lag order	Correlation
China: PPI: Manufacture of Rubber and Plastic Products: MoM	2	CopEnt=-0.401
China: PPI: Manufacture of Automobile: MoM	1	CopEnt=0.434
China: PPI: Manufacture of Automobile: MoM	2	CopEnt=0.367
China: PPI: Manufacture of Automobile: MoM	3	CopEnt=0.532
China: PPI: Manufacture of Automobile: MoM	4	CopEnt=0.492
China: PPI: Manufacture of Automobile: MoM	5	CopEnt=0.561
China: PPI: Manufacture of Automobile: MoM	6	CopEnt=0.634
China: PPI: Manufacture of Automobile: MoM	7	CopEnt=0.599
China: PPI: Manufacture of Automobile: MoM	8	CopEnt=0.862
China: PPI: Manufacture of Automobile: MoM	9	CopEnt=0.738
China: PPI: Manufacture of Automobile: MoM	10	CopEnt=0.712
China: PPI: Manufacture of Automobile: MoM	11	CopEnt=0.818
China: Value of Imports and Exports: CNY	8	Spearman=-0.364
China: Financial Institutions: New RMB Loans	1	Pearson=-0.389
China: Financial Institutions: New RMB Loans: Households: Short-term	11	Pearson=-0.415
China: Financial Institutions: Foreign Currency Loans Balance	12	Spearman=-0.354
China: Financial Institutions: New RMB Deposits	1	Pearson=-0.387
China: Financial Institutions: New RMB Deposits	7	Distance=0.358
Futures Settlement Price (Continuous): Stainless Steel	11	CopEnt=0.771
China: Surveyed Urban Unemployment Rate	1	CopEnt=0.590
China: Manufacturing PMI: Large Enterprises	1	Distance=0.392
China: Manufacturing PMI: Small Enterprises	11	Distance=0.514
帝豪汽车 (Emgrand Automobile)	2	Pearson=0.371
汽车购置税 (Car Purchase Tax)	3	CopEnt=-0.357

Table D.3. Structured feature selection results for the Haval H6 experiment

Feature Name	Lag order	Correlation
China: CPI: Transport and Communications: Transportation Facility: MoM	5	Distance=0.462
China: CPI: Transport and Communications: Transportation Facility: MoM	6	Distance=0.609

(Continued on next page)

(Continued from previous page)

Feature Name	**Lag order**	**Correlation**
China: RPI: MoM	6	Spearman=-0.469
China: RPI: Transportation and Communication Appliances: MoM	6	Distance=0.482
China: PPI: Manufacture of Automobile: MoM	2	CopEnt=0.590
China: PPI: Manufacture of Automobile: MoM	3	CopEnt=0.458
China: PPI: Manufacture of Automobile: MoM	4	CopEnt=0.510
China: PPI: Manufacture of Automobile: MoM	5	CopEnt=0.590
China: PPI: Manufacture of Automobile: MoM	6	CopEnt=0.650
China: PPI: Manufacture of Automobile: MoM	7	CopEnt=0.678
China: PPI: Manufacture of Automobile: MoM	8	CopEnt=0.744
China: PPI: Manufacture of Automobile: MoM	9	CopEnt=0.507
China: PPI: Manufacture of Automobile: MoM	10	CopEnt=0.466
China: PPI: Manufacture of Automobile: MoM	11	CopEnt=0.870
China: Value of Imports and Exports: CNY	8	Spearman=-0.552
China: M1: YoY	2	Distance=0.534
China: Financial Institutions: New RMB Loans	1	Spearman=-0.456
Futures Settlement Price (Active Contract): Stainless Steel	12	CopEnt=0.842
China: Macro-economic Climate Index: Leading Index	1	Pearson=0.459
China: Manufacturing PMI: Small Enterprises	10	Distance=0.556
China: Manufacturing PMI: Small Enterprises	11	Spearman=-0.548
哈弗官网 (Official Website of Haval)	1	Spearman=0.549

Table D.4. Structured feature selection results for the Camry experiment

Feature Name	**Lag order**	**Correlation**
China: PPI: Manufacture of Rubber and Plastic Products: MoM	2	CopEnt=-0.453
China: PPI: Manufacture of Automobile: MoM	2	CopEnt=0.446
China: PPI: Manufacture of Automobile: MoM	3	CopEnt=0.605
China: PPI: Manufacture of Automobile: MoM	4	CopEnt=0.989
China: PPI: Manufacture of Automobile: MoM	5	CopEnt=0.563
China: PPI: Manufacture of Automobile: MoM	6	CopEnt=0.582
China: PPI: Manufacture of Automobile: MoM	7	CopEnt=0.636
China: PPI: Manufacture of Automobile: MoM	8	CopEnt=0.565

(Continued on next page)

(Continued from previous page)

Feature Name	**Lag order**	**Correlation**
China: PPI: Manufacture of Automobile: MoM	9	CopEnt=0.534
China: PPI: Manufacture of Automobile: MoM	10	CopEnt=0.598
China: PPI: Manufacture of Automobile: MoM	11	CopEnt=0.860
China: Export Unit Value Index: HS2: Total Index	12	Distance=0.474
China: Value of Imports and Exports: CNY	12	Spearman=0.734
China: Financial Institutions: New RMB Loans	4	Spearman=0.484
China: Financial Institutions: New RMB Loans	10	Distance=0.441
China: Financial Institutions: New RMB Loans	12	Pearson=0.501
China: Financial Institutions: New RMB Loans: Households	12	Distance=0.597
China: Financial Institutions: New RMB Loans: Households: Short-term	4	Distance=0.408
China: Financial Institutions: New RMB Loans: Households: Short-term	7	Distance=0.404
China: Financial Institutions: New RMB Loans: Households: Short-term	10	Distance=0.485
Futures Settlement Price (Continuous): Deformed Steel Bar	12	Distance=0.711
Futures Settlement Price (Continuous): Stainless Steel	12	CopEnt=0.933
China: Manufacturing PMI: Small Enterprises	8	Spearman=0.417
China: Manufacturing PMI: Small Enterprises	9	Distance=0.405
China: Manufacturing PMI: Small Enterprises	12	Spearman=0.618
凯美瑞油耗 (Fuel Consumption of Camry)	3	Distance=0.436
凯美瑞油耗 (Fuel Consumption of Camry)	12	Distance=0.440
丰田凯美瑞 (Toyota Camry)	7	Distance=0.420
汽车保险 (Car Insurance)	5	Spearman=-0.723
汽车摇号 (Car Lottery)	8	Spearman=-0.595
汽车摇号 (Car Lottery)	12	Distance=0.478
汽车销量排行 (Car Sales Ranking)	2	Spearman=-0.451
汽车销量排行 (Car Sales Ranking)	4	Spearman=-0.522
汽车销量排行 (Car Sales Ranking)	6	Spearman=-0.529
汽车销量排行 (Car Sales Ranking)	8	Spearman=-0.485
汽车销量排行 (Car Sales Ranking)	10	Spearman=-0.494

Appendix E. Description of benchmark methods

In our experiments, the benchmark methods used for comparison include the (S, s), DQN, A2C, SAC, PPO, and DDPG. Detailed descriptions are as follows:

(1) (S, s) Policy (Arrow et al., 1958): This is a classical periodic inventory control strategy. The core rule is: at the end of each period, the inventory level is checked; if the current inventory is less than or equal to threshold s (reorder point), a replenishment is triggered to raise the inventory to the target level S; otherwise, no replenishment is performed. S and s are determined based on historical sales statistics such as the mean and standard deviation, where s is set to the historical mean demand and S is calculated as:

$$S = \mu + z \times \sigma,\ z = \Phi^{-1}\left(\frac{c_b}{c_b + c_h}\right), \tag{E.1}$$

where Φ^{-1} is the inverse of the standard normal distribution, and μ and σ represent the estimated mean and standard deviation of demand based on historical sales.

(2) DQN (Deep Q-Network) (Oroojlooyjadid et al., 2021): This method integrates deep neural networks with reinforcement learning. Its core idea is to approximate the Q-value function (i.e., state-action value function) using a neural network, thereby learning the optimal policy in complex environments. In the multi-period inventory context, DQN defines a discrete action space by evenly dividing the range between the minimum and maximum historical demand values (e.g., 10,000 discrete replenishment quantities). The policy is optimized by maximizing cumulative rewards (i.e., minimizing total costs). During training, the ε-greedy strategy balances exploration and exploitation, selecting a random action with certain probability or otherwise choosing the action with the highest current Q-value. Network parameters are updated by minimizing the temporal-difference (TD) error, and the target Q-values are calculated using a separately updated target network to stabilize training.

(3) A2C (Advantage Actor-Critic) (Mohamadi et al., 2024): A policy gradient-based on-policy reinforcement learning algorithm that employs a shared network architecture to jointly optimize the policy function and value function. The Actor network outputs the parameters of a Gaussian distribution (mean μ and standard deviation σ) over the continuous action space for replenishment, with reparameterization techniques used for differentiable sampling. The Critic network evaluates the state value function $V(\mathbf{S})$, and shares lower layers with the Actor to improve training efficiency. A2C uses Generalized Advantage Estimation (GAE), $A_t^{GAE} = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}$, which balances bias and variance by incorporating multi-step returns. An entropy regularization term $H(\pi(\cdot|\mathbf{S}))$ is added to prevent premature convergence to local optima. Gradient clipping and advantage normalization are also applied. The optimization objective is: $J(\theta) = \mathbb{E}[\log \pi(a|\mathbf{S})A_t + \beta H(\pi)]$. This approach learns adaptive replenishment policies under stochastic demand, and the shared network architecture improves training efficiency.

(4) SAC (Soft Actor-Critic) (Kou et al., 2025): This is an off-policy algorithm based on the maximum entropy reinforcement learning framework. Its key feature is the dual objec-

tive of maximizing both expected returns and policy entropy. Unlike traditional Actor-Critic methods such as A2C, SAC employs a stochastic policy instead of a deterministic one, encouraging exploration through entropy regularization. The Actor network outputs a probability distribution (e.g., Gaussian) from which continuous replenishment actions are sampled, overcoming the granularity limitations of discretization in DQN. Two independent Critic networks (Q-functions) and a temperature parameter α are used to dynamically balance exploration and exploitation. In inventory management applications, SAC's entropy-maximization design enhances its adaptability to demand fluctuations and non-stationary environments. The SAC objective is expressed as: $J(\pi) = \mathbb{E}[\Sigma(r(\mathbf{S}, a) + \alpha H(\pi(\cdot|\mathbf{S})))]$, where $H(\pi)$ denotes policy entropy and α automatically adjusts the degree of exploration.

(5) PPO (Proximal Policy Optimization) (Schulman et al., 2017): This method strikes a balance between training stability and decision accuracy via constrained policy updates and an adaptive optimization framework. The core innovation lies in its clipped surrogate objective function: $L^{CLIP}(\theta) = \mathbb{E}\left[\min\left(r(\theta), \text{clip}\left(r(\theta), 1-\epsilon, 1+\epsilon\right)\right) A(\mathbf{S}, a)\right]$, where the policy ratio $r(\theta) = \pi_\theta(a|\mathbf{S})/\pi_{\theta_{\text{old}}}(a|\mathbf{S})$ quantifies the extent of policy update, and the clipping threshold ϵ (typically 0.1-0.3) constrains deviation to prevent policy oscillation, particularly in inventory settings without lead time, where abrupt changes in replenishment could destabilize the system. PPO applies multiple mini-batch updates (3-10 iterations per batch) and state-related constraints to stably extract policy gradients from limited-period observations. Its entropy regularization term $H(\pi_\theta) = -\mathbb{E}[\Sigma_a \pi_\theta(a|\mathbf{S}) \log \pi_\theta(a|\mathbf{S})]$ dynamically adjusts exploration intensity to avoid premature convergence under non-stationary demand. While PPO sacrifices some historical data efficiency compared to off-policy methods like SAC, its hard constraint on policy update magnitude reduces the risk of policy collapse. In practice, the clipping threshold ϵ and learning rate should be tuned based on the coefficient of demand variation: $\epsilon = 0.1$ is recommended for high volatility, while 0.3 is suitable for stable demand to accelerate convergence.

(6) DDPG (Deep Deterministic Policy Gradient) (Lillicrap et al., 2016): This off-policy algorithm combines deep neural networks with deterministic policy gradients, directly generating continuous replenishment actions via an Actor-Critic architecture. It consists of a deterministic Actor network and a Critic network for Q-value estimation. Unlike DQN, which operates on a discrete action space, DDPG supports continuous action outputs, eliminating discretization bias. Compared with SAC's stochastic policy, DDPG's deterministic policy is more efficient under stable demand conditions. As the foundation of the TD3 algorithm used in this study, DDPG exhibits certain limitations in continuous action space control: it uses only a single Critic network, making its value estimation prone to overestimation. Additionally, its simultaneous updates of the Actor and Critic networks introduce instability, and it lacks the target policy smoothing regularization employed by TD3. In inventory management scenarios, DDPG demonstrates lower convergence speed and less training stability than TD3.

Appendix F. Hyperparameter search space and selected results

Similar to De Moor et al. (2022), for each method and each hyperparameter in the experiments, a finite set of candidate values was predefined. The optimal combination was then selected based on performance on the validation set. It is worth noting that due to computational limitations, an exhaustive search over all possible combinations was infeasible. Therefore, we adopted a random grid search strategy. The results reported are relatively conservative, and more extensive hyperparameter tuning could potentially further improve the performance of the proposed method.

(1) *Hyperparameter selection for WET-TD3*

The hyperparameters involved in WET-TD3, along with their symbols, meanings, and search spaces, are listed in Table F.1. After repeated experimentation, we determine that across all product experiments and (c_h, c_b) combinations, the optimal values are as follows: $\gamma = 0.83$; $\mu = 0.01$; $\sigma = 0.4$; $\tilde{\sigma} = 0.6$; $\varsigma_0 = 0.5$; Delay = 2; $Ep_{\max} = 500$; $T_{\max} = 100$; $B_{\max} = 200$; $\Delta_{\min} = 0.001$. The values of other hyperparameters under different product and (c_h, c_b) combinations are shown in Table F.2.

Table F.1. Introduction to the hyperparameters of WET-TD3.

Hyperparameter	Symbol	Description	Search Range
Embedding Dim	v	Dimension of the trainable weight matrix $\mathbf{W}_E$ in the embedding layer	8, 16, 20, 24, 28, 32, 36, 40, 64
Encoder Layers	L_E	Number of layers in Transformer encoder	1, 2, 3, 4
Feedforward Dim	d_E	Hidden layer dimension of FFN in Transformer encoder	16, 32, 64, 128, 256, 512, 1024
Discount	γ	Discount factor	0.80, 0.81, 0.82, ..., 0.99
Update Rate	μ	Soft update rate of target networks	0.005, 0.01, 0.05, 0.1
Action noise	σ	Standard deviation of Gaussian noise ε on action a	0.2, 0.4, 0.6, 0.8
Policy noise	$\tilde{\sigma}$	Std. of policy noise ς	0.2, 0.4, 0.6, 0.8
Policy noise clip	ς_0	Clipping value for policy noise	0.5, 0.6, 0.7, 0.8
Delay	delay	Delay steps for target network updates	1, 2, 3, 4
Max Episodes	$Ep_{\max}$	Maximum training episodes	500, 1000
Max Timestep	$T_{\max}$	Maximum time steps per episode	100, 200
Max Size	$B_{\max}$	Size of replay buffer $\mathcal{B}$	200, 300, 400
Batch Size	N	Batch size sampled per update	8, 16, 32, 64

(Continued on next page)

(Continued from previous page)

Hyperparameter	**Symbol**	**Description**	**Search Range**
Learning Rate A	lr_A	Learning rate for Actor and target Actor	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Learning Rate Q	lr_Q	Learning rate for Critic and target Critic	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Min Improv	$\mathit{\Delta}_{\min}$	Minimum loss improvement for early stopping	0.01, 0.001, 0.0001
Patience	P	Epochs to wait without improvement	5, 10, 15, 20, ..., 150

Table F.2. Selected hyperparameters for training WET-TD3.

Experiment	(c_h, c_b)	v	L_E	d_E	N	lr_A	lr_Q	P
Lavida	(1,0.5)	24	4	16	8	1E-03	1E-02	70
	(1,1)	24	4	16	8	1E-03	1E-02	100
	(1,2)	20	1	64	64	1E-05	1E-02	60
	(1,5)	20	2	64	64	1E-05	1E-02	70
	(1,10)	20	1	64	64	1E-05	1E-02	70
	(1,20)	20	1	64	64	1E-05	1E-02	70
	(1,50)	20	1	128	64	1E-05	1E-02	100
	(1,100)	20	1	64	64	1E-05	1E-02	50
Emgrand	(1,0.5)	24	4	16	8	1E-03	1E-02	100
	(1,1)	24	1	1024	32	1E-03	1E-02	20
	(1,2)	24	1	512	32	1E-05	1E-02	20
	(1,5)	24	1	64	64	1E-05	1E-02	100
	(1,10)	24	1	64	64	1E-05	1E-02	100
	(1,20)	24	1	512	32	1E-05	1E-02	20
	(1,50)	24	1	512	64	1E-05	1E-02	50
	(1,100)	24	1	256	32	1E-03	1E-02	50
Haval H6	(1,0.5)	20	1	64	16	1E-03	1E-02	20
	(1,1)	20	1	64	64	1E-03	1E-02	10
	(1,2)	16	1	16	64	1E-03	1E-02	30
	(1,5)	20	1	128	8	1E-03	1E-02	100
	(1,10)	16	1	16	16	1E-03	1E-02	10
	(1,20)	16	1	16	16	1E-03	1E-02	100
	(1,50)	16	1	16	8	1E-03	1E-02	135
	(1,100)	16	1	16	8	1E-03	1E-02	30

(Continued on next page)

(Continued from previous page)

Experiment	(c_h, c_b)	v	L_E	d_E	N	lr_A	lr_Q	P
Camry	(1,0.5)	16	1	16	8	1E-03	1E-02	10
	(1,1)	24	1	128	32	1E-03	1E-02	10
	(1,2)	24	1	64	32	1E-03	1E-02	20
	(1,5)	16	1	16	32	1E-03	1E-02	20
	(1,10)	16	1	64	32	1E-03	1E-02	20
	(1,20)	16	1	64	32	1E-04	1E-02	10
	(1,50)	16	1	16	64	1E-04	1E-03	50
	(1,100)	20	1	64	64	1E-04	1E-01	10

(2) *Hyperparameter selection for No_Feat*

The hyperparameters involved in No_Feat, along with their symbols, meanings, and search ranges, are the same as in Table F.1. After repeated experiments, the optimal values for all product experiments and (c_h, c_b) combinations are: $\gamma = 0.83$; $\mu = 0.01$; $\sigma = 0.4$; $\tilde{\sigma} = 0.6$; $\varsigma_0 = 0.5$; Delay $= 2$; $Ep_{\max} = 500$; $T_{\max} = 100$; $B_{\max} = 200$; $\varDelta_{\min} = 0.001$. The values of other hyperparameters under different products and (c_h, c_b) combinations are shown in Table F.3.

Table F.3. Selected hyperparameters for training No_Feat.

Experiment	(c_h, c_b)	v	L_E	d_E	N	lr_A	lr_Q	P
Lavida	(1,0.5)	24	4	16	8	1E-03	1E-02	100
	(1,1)	24	4	16	8	1E-03	1E-02	100
	(1,2)	20	1	64	64	1E-05	1E-02	100
	(1,5)	20	2	64	64	1E-05	1E-02	100
	(1,10)	20	1	64	64	1E-05	1E-02	100
	(1,20)	20	1	64	64	1E-05	1E-02	100
	(1,50)	20	1	64	64	1E-05	1E-02	100
	(1,100)	20	1	64	64	1E-04	1E-02	100
Emgrand	(1,0.5)	24	4	16	8	1E-03	1E-02	100
	(1,1)	8	1	64	8	1E-03	1E-02	100
	(1,2)	24	1	256	64	1E-05	1E-02	100
	(1,5)	24	1	64	64	1E-05	1E-02	100
	(1,10)	24	1	64	64	1E-05	1E-02	100
	(1,20)	24	1	64	64	1E-05	1E-02	100
	(1,50)	16	1	64	64	1E-05	1E-02	100
	(1,100)	16	1	64	64	1E-05	1E-02	100
Haval H6	(1,0.5)	20	1	16	8	1E-03	1E-02	100
	(1,1)	20	1	64	64	1E-03	1E-02	100

(Continued on next page)

Experiment	(c_h, c_b)	v	L_E	d_E	N	lr_A	lr_Q	P
	(1,2)	16	1	64	8	1E-03	1E-02	100
	(1,5)	24	1	128	16	1E-04	1E-02	200
	(1,10)	16	1	64	8	1E-03	1E-02	100
	(1,20)	24	1	64	8	1E-05	1E-02	100
	(1,50)	24	1	64	8	1E-05	1E-02	100
	(1,100)	24	3	64	64	1E-05	1E-02	100
Camry	(1,0.5)	24	4	16	8	1E-03	1E-02	100
	(1,1)	8	1	64	8	1E-03	1E-02	100
	(1,2)	16	1	64	64	1E-03	1E-02	20
	(1,5)	24	1	64	64	1E-05	1E-02	100
	(1,10)	24	1	64	64	1E-05	1E-02	100
	(1,20)	24	1	64	64	1E-04	1E-02	10
	(1,50)	16	1	64	64	1E-04	1E-03	20
	(1,100)	20	1	64	64	1E-04	1E-01	10

(3) *Hyperparameter selection for No_Rev*

The hyperparameters involved in No_Rev, along with their symbols, meanings, and search ranges, are the same as in Table F.1, except that it does not include the hyperparameters Embedding Dim, Encoder Layers, and Feedforward Dim. After repeated experiments, the optimal values across all product experiments and (c_h, c_b) combinations are: $\mu = 0.01$; $\sigma = 0.4$; $\tilde{\sigma} = 0.6$; $\varsigma_0 = 0.5$; Delay = 2; $Ep_{\max} = 500$; $T_{\max} = 100$; $B_{\max} = 200$; $\mathit{\Delta}_{\min} = 0.001$. Other hyperparameter values are shown in Table F.4.

Table F.4. Selected hyperparameters for training No_Rev.

Experiment	(c_h, c_b)	γ	N	lr_A	lr_Q	P
Lavida	(1,0.5)	0.83	64	1E-04	5E-02	100
	(1,1)	0.83	64	1E-03	1E-03	100
	(1,2)	0.83	64	1E-03	5E-02	100
	(1,5)	0.92	64	1E-03	1E-04	100
	(1,10)	0.83	64	1E-05	1E-03	100
	(1,20)	0.92	64	1E-05	1E-02	100
	(1,50)	0.83	64	1E-05	1E-02	50
	(1,100)	0.83	64	1E-04	1E-02	100
Emgrand	(1,0.5)	0.83	64	1E-03	5E-02	100
	(1,1)	0.83	64	1E-05	1E-02	100
	(1,2)	0.83	64	1E-04	1E-02	100
	(1,5)	0.83	64	1E-03	1E-02	100

(Continued on next page)

Experiment	(c_h, c_b)	γ	N	lr_A	lr_Q	P
	(1,10)	0.83	64	1E-05	1E-03	100
	(1,20)	0.92	64	1E-05	1E-02	100
	(1,50)	0.83	64	1E-05	1E-02	100
	(1,100)	0.83	64	1E-05	1E-02	100
Haval H6	(1,0.5)	0.83	64	1E-03	5E-02	100
	(1,1)	0.83	64	1E-03	1E-03	100
	(1,2)	0.83	64	1E-03	5E-02	100
	(1,5)	0.83	64	1E-03	1E-03	100
	(1,10)	0.83	64	1E-05	1E-03	100
	(1,20)	0.83	64	1E-03	1E-03	100
	(1,50)	0.83	64	1E-05	1E-03	100
	(1,100)	0.83	64	1E-05	1E-03	100
Camry	(1,0.5)	0.83	64	1E-03	5E-02	100
	(1,1)	0.83	64	1E-05	1E-02	100
	(1,2)	0.83	64	1E-04	1E-02	100
	(1,5)	0.83	64	1E-03	1E-02	100
	(1,10)	0.83	32	1E-03	1E-02	50
	(1,20)	0.83	64	1E-02	1E-01	50
	(1,50)	0.83	8	1E-02	1E-01	50
	(1,100)	0.83	8	1E-02	1E-01	50

(4) *Hyperparameter selection for DDPG*

The hyperparameters involved in DDPG, along with their symbols, meanings, and search ranges, are the same as in Table F.1, except that it does not include the hyperparameters Embedding Dim, Encoder Layers, Feedforward Dim, and delay. After repeated experiments, the optimal values across all product experiments and (c_h, c_b) combinations are: $\mu = 0.01$; $\sigma = 0.4$; $\tilde{\sigma} = 0.6$; $\varsigma_0 = 0.5$; $Ep_{\max} = 500$; $T_{\max} = 100$; $B_{\max} = 200$; $N = 64$; $\mathit{\Delta}_{\min} = 0.001$. Other hyperparameter values are shown in Table F.5.

Table F.5. Selected hyperparameters for training DDPG.

Experiment	(c_h, c_b)	γ	lr_A	lr_Q	P
Lavida	(1,0.5)	0.83	1E-03	1E-04	50
	(1,1)	0.83	1E-03	1E-04	50
	(1,2)	0.83	1E-03	1E-03	50
	(1,5)	0.92	1E-03	1E-03	50
	(1,10)	0.83	1E-03	1E-03	50
	(1,20)	0.83	1E-03	1E-03	50

(Continued on next page)

Experiment	(c_h, c_b)	γ	lr_A	lr_Q	P
	(1,50)	0.83	1E-03	1E-03	50
	(1,100)	0.83	1E-03	1E-03	50
Emgrand	(1,0.5)	0.83	1E-03	5E-02	100
	(1,1)	0.83	1E-05	1E-02	100
	(1,2)	0.83	1E-03	1E-02	50
	(1,5)	0.83	1E-04	1E-02	50
	(1,10)	0.83	1E-05	1E-03	100
	(1,20)	0.92	1E-05	1E-02	100
	(1,50)	0.83	1E-05	1E-03	50
	(1,100)	0.83	1E-04	1E-03	50
Haval H6	(1,0.5)	0.83	1E-03	5E-02	100
	(1,1)	0.83	1E-03	1E-03	100
	(1,2)	0.83	1E-03	5E-02	100
	(1,5)	0.83	1E-03	1E-03	100
	(1,10)	0.83	1E-05	1E-03	100
	(1,20)	0.83	1E-03	1E-03	100
	(1,50)	0.83	1E-05	1E-03	100
	(1,100)	0.83	1E-05	1E-03	100
Camry	(1,0.5)	0.83	1E-03	5E-02	100
	(1,1)	0.83	1E-03	1E-03	100
	(1,2)	0.83	1E-03	5E-02	100
	(1,5)	0.83	1E-03	1E-03	100
	(1,10)	0.83	1E-03	1E-02	50
	(1,20)	0.83	5E-03	1E-02	50
	(1,50)	0.83	1E-02	1E-01	50
	(1,100)	0.83	1E-02	1E-01	50

(5) *Hyperparameter selection for EAS_Same*

The hyperparameters involved in EAS_Same, along with their symbols, meanings, and search ranges, are listed in Table F.6. After repeated experiments, the optimal values across all product experiments are: $v = 16$; $d_E = 64$; $Ep_{\max} = 500$; $N = 64$; $\varDelta_{\min} = 0.001$. Other values are shown in Table F.7.

Table F.6. Introduction to the hyperparameters of EAS_Same

Hyperparameter Name	Symbol	Meaning	Search Space
Embedding Dim	v	Dimension of the trainable weight matrix $\mathbf{W}_E$ in the word embedding layer	8, 16, 20, 24, 28, 32, 36, 40, 64
Encoder Layers	L_E	Number of Transformer encoder layers	1, 2, 3, 4
Feedforward Dim	d_E	Dimension of the hidden layer in the Transformer encoder’s feedforward network	16, 32, 64, 128, 256, 512, 1024
Max Episodes	$Ep_{\max}$	Maximum number of training episodes	500, 1000
Batch Size	N	Batch size per iteration	8, 16, 32, 64
Learning Rate	lr	Learning rate	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Min Improv	$\varDelta_{\min}$	Minimum improvement in loss for early stopping	0.01, 0.001, 0.0001
Patience	P	Number of epochs allowed without validation improvement	5, 10, 15, 20, ..., 150

Table F.7. Selected hyperparameters for training EAS_Same.

Experiment	L_E	lr	P
Lavida	2	0.01	100
Emgrand	2	0.01	100
Haval H6	1	0.001	100
Camry	1	0.01	50

(6) *Hyperparameter selection for DQN*

The hyperparameters involved in DQN, along with their symbols, meanings, and search ranges, are listed in Table F.8. After repeated experiments, the optimal values across all product experiments and (c_h, c_b) combinations are: $Ep_{\max} = 500$; $B_{\max} = 200$; $N = 64$; $\varDelta_{\min} = 0.001$; $P = 50$. Other values are shown in Table F.9.

Table D.8. Introduction to the hyperparameters of DQN.

Hyperparameter Name	Symbol	Meaning	Search Space
Discount	γ	Discount factor	0.80, 0.81, 0.82, ..., 0.99
Epsilon	ε	Exploration probability in ε-greedy policy	0.001, 0.005, 0.01, 0.05, 0.1
Delay	delay	Steps before target network update	1, 2, 3, 4
Max Episodes	$Ep_{\max}$	Maximum training episodes	500, 1000
Max Size	$B_{\max}$	Size of replay buffer $\mathcal{B}$	200, 300, 400
Batch Size	N	Sample batch size from experience buffer	8, 16, 32, 64
Learning Rate	lr	Learning rate	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Min Improv	$\varDelta_{\min}$	Minimum loss improvement for early stopping	0.01, 0.001, 0.0001
Patience	P	Epochs allowed without validation improvement	5, 10, 15, 20, ..., 150

Table D.9. Selected hyperparameters for training DQN.

Experiment	(c_h, c_b)	γ	ε	**delay**	lr
Lavida	(1,0.5)	0.83	0.01	2	1E-03
	(1,1)	0.83	0.01	2	1E-03
	(1,2)	0.83	0.01	2	1E-03
	(1,5)	0.83	0.01	2	1E-03
	(1,10)	0.83	0.01	2	1E-03
	(1,20)	0.83	0.01	2	1E-02
	(1,50)	0.83	0.01	2	1E-03
	(1,100)	0.83	0.01	2	1E-03
Emgrand	(1,0.5)	0.83	0.01	2	1E-03
	(1,1)	0.83	0.01	2	1E-03
	(1,2)	0.83	0.01	2	1E-03
	(1,5)	0.90	0.01	2	1E-03
	(1,10)	0.83	0.01	2	1E-03

(Continued on next page)

Experiment	(c_h, c_b)	γ	ε	delay	lr
	(1,20)	0.92	0.01	2	1E-03
	(1,50)	0.83	0.01	2	1E-03
	(1,100)	0.83	0.01	2	1E-03
Haval H6	(1,0.5)	0.83	0.01	2	1E-02
	(1,1)	0.83	0.01	2	1E-02
	(1,2)	0.80	0.01	2	1E-02
	(1,5)	0.95	0.01	2	1E-03
	(1,10)	0.83	0.01	2	5E-03
	(1,20)	0.95	0.01	2	1E-02
	(1,50)	0.90	0.01	2	1E-03
	(1,100)	0.90	0.01	2	1E-03
Camry	(1,0.5)	0.83	0.01	2	1E-03
	(1,1)	0.83	0.01	2	1E-03
	(1,2)	0.83	0.01	2	1E-03
	(1,5)	0.83	0.01	2	1E-03
	(1,10)	0.99	0.01	2	1E-03
	(1,20)	0.99	0.001	3	1E-01
	(1,50)	0.99	0.001	4	1E-01
	(1,100)	0.90	0.001	2	1E-01

(7) *Hyperparameter selection for A2C*

The hyperparameters involved in A2C, along with their symbols, meanings, and search ranges, are listed in Table F.10. After repeated experiments, the optimal values across all product experiments and (c_h, c_b) combinations are: $Ep_{\max} = 500$; $T_{\max} = 200$; $\mathit{\Delta}_{\min} = 0.001$; $P = 50$. Other values are shown in Table F.11.

Table D.10. Introduction to the hyperparameters of A2C.

Hyperparameter Name	Symbol	Meaning	Search Space
Discount	γ	Discount factor	0.80, 0.81, 0.82, ..., 0.99
Entropy Coefficient	β	Entropy regularization to encourage exploration	0.001, 0.01, 0.1
Max Episodes	$Ep_{\max}$	Maximum training episodes	500, 1000
GAE Lambda	λ	GAE bias-variance tradeoff parameter	0.5, 1.0, 1.5, ..., 0.95, 0.99

(Continued on next page)

Hyperparameter Name	Symbol	Meaning	Search Space
Batch Size	N	Training batch size per iteration	8, 16, 32, 64
Learning Rate	lr	Learning rate	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Max Timestep	$T_{\max}$	Maximum training steps per episode	100, 200
Patience	P	Epochs allowed without validation improvement	5, 10, 15, 20, ..., 150

Table D.11. Selected hyperparameters for training A2C.

Experiment	(c_h, c_b)	γ	β	lr	λ	N
Lavida	(1,0.5)	0.99	0.01	1E-02	0.95	32
	(1,1)	0.99	0.01	1E-02	0.95	32
	(1,2)	0.99	0.01	1E-02	0.95	32
	(1,5)	0.99	0.01	1E-02	0.95	32
	(1,10)	0.99	0.01	1E-01	0.90	64
	(1,20)	0.99	0.01	5E-02	0.80	64
	(1,50)	0.99	0.001	1E-01	0.70	64
	(1,100)	0.99	0.01	5E-02	0.83	64
Emgrand	(1,0.5)	0.90	0.01	5E-03	0.95	64
	(1,1)	0.90	0.01	5E-03	0.95	64
	(1,2)	0.90	0.01	5E-03	0.95	64
	(1,5)	0.83	0.01	1E-02	0.95	64
	(1,10)	0.90	0.001	1E-02	0.99	64
	(1,20)	0.90	0.01	1E-01	0.80	64
	(1,50)	0.95	0.001	5E-02	0.85	64
	(1,100)	0.90	0.001	1E-01	0.80	64
Haval H6	(1,0.5)	0.83	0.01	1E-02	0.95	32
	(1,1)	0.83	0.01	1E-02	0.95	32
	(1,2)	0.99	0.01	1E-03	0.99	64
	(1,5)	0.99	0.01	1E-02	0.95	64
	(1,10)	0.80	0.001	1E-01	0.95	64
	(1,20)	0.99	0.01	1E-01	0.85	64
	(1,50)	0.85	0.001	1E-01	0.80	64

(Continued on next page)

Experiment	(c_h, c_b)	γ	β	lr	λ	N
	(1,100)	0.99	0.001	1E-02	0.95	64
Camry	(1,0.5)	0.83	0.01	1E-02	0.95	64
	(1,1)	0.83	0.01	1E-02	0.99	64
	(1,2)	0.99	0.01	1E-02	0.99	64
	(1,5)	0.95	0.001	5E-02	0.85	64
	(1,10)	0.99	0.001	5E-02	0.99	64
	(1,20)	0.99	0.001	5E-02	0.99	64
	(1,50)	0.99	0.01	1E-01	0.85	64
	(1,100)	0.97	0.01	5E-02	0.50	64

(8) *Hyperparameter selection for SAC*

The hyperparameters involved in SAC, along with their symbols, meanings, and search ranges, are listed in Table F.12. After repeated experiments, the optimal values across all product experiments and (c_h, c_b) combinations are: $\gamma = 0.83$; $\mu = 0.01$; $Ep_{\max} = 500$; $T_{\max} = 100$; $B_{\max} = 200$; $\Delta_{\min} = 0.001$. Other values are shown in Table F.13.

Table F.12. Introduction to the hyperparameters of SAC.

Hyperparameter Name	**Symbol**	**Meaning**	**Search Space**
Discount	γ	Discount factor	0.0.80, 0.81, 0.82, ..., 0.99
Update Rate	μ	Target network soft update rate	0.001, 0.005, 0.01, 0.05, 0.1
Max Episodes	$Ep_{\max}$	Maximum training episodes	500, 1000
Max Timestep	$T_{\max}$	Maximum steps per episode	100, 200
Max Size	$B_{\max}$	Replay buffer size $\mathcal{B}$	200, 300, 400
Batch Size	N	Batch size per iteration	8, 16, 32, 64
Learning Rate A	lr_A	Actor and target Actor learning rate	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Learning Rate Q	lr_Q	Critic and target Critic learning rate	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Min Improv	$\Delta_{\min}$	Min loss improvement for early stop	0.01, 0.001, 0.0001
Patience	P	Epochs allowed without validation improvement	5, 10, 15, 20, ..., 150

Table F.13. Selected hyperparameters for training SAC.

Experiment	(c_h, c_b)	N	lr_A	lr_Q	P
Lavida	(1,0.5)	64	1E-04	1E-03	50
	(1,1)	64	1E-04	1E-03	20
	(1,2)	64	1E-04	1E-02	50
	(1,5)	64	1E-04	1E-04	50
	(1,10)	64	1E-04	1E-03	50
	(1,20)	64	1E-05	1E-02	100
	(1,50)	64	1E-05	1E-03	100
	(1,100)	64	1E-03	1E-02	50
Emgrand	(1,0.5)	32	1E-04	1E-03	50
	(1,1)	64	1E-05	1E-02	100
	(1,2)	32	1E-05	1E-03	50
	(1,5)	64	1E-05	1E-03	50
	(1,10)	64	1E-05	1E-03	50
	(1,20)	64	1E-05	1E-03	50
	(1,50)	64	1E-04	1E-02	50
	(1,100)	64	1E-05	1E-03	100
Haval H6	(1,0.5)	64	1E-03	5E-02	100
	(1,1)	64	1E-03	1E-03	100
	(1,2)	64	1E-03	1E-02	100
	(1,5)	64	1E-03	1E-03	100
	(1,10)	64	1E-05	1E-03	100
	(1,20)	64	1E-03	1E-03	100
	(1,50)	64	1E-05	1E-03	100
	(1,100)	64	1E-05	1E-03	100
Camry	(1,0.5)	64	1E-03	5E-02	100
	(1,1)	64	1E-05	1E-02	100
	(1,2)	64	1E-04	1E-02	100
	(1,5)	64	1E-03	1E-02	100
	(1,10)	32	1E-03	1E-02	100
	(1,20)	64	1E-02	1E-01	100
	(1,50)	32	1E-02	1E-02	50
	(1,100)	32	1E-02	1E-02	50

(9) *Hyperparameter selection for PPO*

The hyperparameters involved in PPO, along with their symbols, meanings, and search ranges, are listed in Table F.14. After repeated experiments, the optimal values across all product experiments and (c_h, c_b) combinations are: $\epsilon = 0.2$; $\gamma = 0.99$; $Ep_{\max} = 500$; $B_{\max} = 200$; $N = 64$; $\varDelta_{\min} = 0.001$; $P = 50$. Other values are shown in Table F.15.

Table F.14. Introduction to the hyperparameters of PPO.

Hyperparameter Name	**Symbol**	**Meaning**	**Search Space**
Discount	γ	Discount factor	0.80, 0.81, 0.82, ..., 0.99
Clip Epsilon	ϵ	Clipping threshold for policy update	0.001, 0.01, 0.1
Max Episodes	$Ep_{\max}$	Maximum training episodes	500, 1000
GAE Lambda	λ	GAE bias-variance parameter	0.9, 0.95, 0.99
Max Size	$B_{\max}$	Replay buffer size	200, 300, 400
Batch Size	N	Training batch size per iteration	8, 16, 32, 64
Learning Rate A	lr_A	Policy network learning rate	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Learning Rate Q	lr_Q	Value network learning rate	1E-1, 1E-2, 5E-2, 1E-3, 5E-3, 1E-4, 5E-4, 1E-5
Update Epochs	Ep_u	Policy update epochs per batch	10, 20, 30, 40, 50
Max Timestep	$T_{\max}$	Max steps per episode	100, 200
Patience	P	Epochs allowed without validation improvement	5, 10, 15, 20, ..., 150

Table F.15. Selected hyperparameters for training PPO.

Experiment	(c_h, c_b)	λ	lr_A	lr_Q	Ep_u
Lavida	(1,0.5)	0.95	1E-03	1E-03	10
	(1,1)	0.95	1E-03	1E-03	20
	(1,2)	0.95	1E-03	1E-03	40
	(1,5)	0.90	1E-03	1E-03	20
	(1,10)	0.95	1E-03	1E-03	20
	(1,20)	0.99	1E-03	1E-03	40
	(1,50)	0.95	1E-03	1E-03	20

(Continued on next page)

Experiment	(c_h, c_b)	λ	lr_A	lr_Q	Ep_u
	(1,100)	0.99	1E-03	1E-03	40
Emgrand	(1,0.5)	0.99	1E-03	1E-03	20
	(1,1)	0.95	1E-03	1E-03	20
	(1,2)	0.99	1E-03	1E-03	20
	(1,5)	0.99	1E-03	1E-03	20
	(1,10)	0.99	1E-03	1E-03	20
	(1,20)	0.99	1E-03	1E-03	30
	(1,50)	0.99	1E-03	5E-03	30
	(1,100)	0.99	1E-03	5E-03	40
Haval H6	(1,0.5)	0.99	1E-03	1E-03	20
	(1,1)	0.99	1E-03	1E-03	20
	(1,2)	0.95	1E-03	1E-03	40
	(1,5)	0.99	1E-03	1E-03	30
	(1,10)	0.99	1E-03	1E-03	30
	(1,20)	0.95	1E-03	1E-03	10
	(1,50)	0.95	1E-03	1E-03	30
	(1,100)	0.99	1E-03	1E-03	20
Camry	(1,0.5)	0.95	1E-03	1E-03	20
	(1,1)	0.95	1E-03	1E-03	20
	(1,2)	0.95	1E-03	1E-03	20
	(1,5)	0.99	1E-03	1E-03	40
	(1,10)	0.99	5E-03	5E-03	20
	(1,20)	0.99	1E-03	1E-03	30
	(1,50)	0.95	5E-03	1E-02	40
	(1,100)	0.95	1E-01	1E-01	20

References

Arrow, K.J., Karlin, S., & Scarf, H. (1958). Studies in the Mathematical Theory of Inventory and Production. Stanford University Press.

De Moor, B.J., Gijsbrechts, J., & Boute, R.N. (2022). Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management. *European Journal of Operational Research*, 301(2), 535-545.

Kou, A., Cheng, Y., Huang, X., & Jin, J. (2025). Dynamic replenishment policy for perishable goods using change point detection-based soft actor-critic reinforcement learning. *Expert Systems with Applications*, 270, 126556.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). Continuous control with deep reinforcement learning. 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, San Juan.

Ma, J. (2021). Variable Selection with Copula Entropy. *Chinese Journal of Applied Probability and Statistics*, 34(7), 405–420.

Mohamadi, N., Niaki, S.T.A., Taher, M., & Shavandi, A. (2024). An application of deep reinforcement learning and vendor-managed inventory in perishable supply chain management. *Engineering Applications of Artificial Intelligence*, 127, 107403.

Myers, J., Well, A., & Lorch, R. (2010). Research Design and Statistical Analysis, Third Edition. Routledge.

Oroojlooyjadid, A., Nazari, M., Snyder, L.V., & Takáč, M. (2021). A deep Q-network for the beer game: Deep reinforcement learning for inventory optimization. *Manufacturing & Service Operations Management*, 24(1), 285-304.

Rodgers, J.L., & Nicewander, W.A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1), 59-66.

Schnaubelt, M. (2022). Deep reinforcement learning for the optimal placement of cryptocurrency limit orders. *European Journal of Operational Research*, 296(3), 993-1006.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint*, Article arXiv:1707.06347.

Székely, G.J., Rizzo, M.L., & Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6), 2769-2794.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, NIPS 2017, Long Beach.