D ADDITIONAL EXPERIMENTS

To evaluate the effectiveness of our proposed TS-Reasoner, we compared it with state-of-the-art reasoning-based models to highlight its advantages in decision-making, compositional optimization, and causal reasoning scenarios. These experiments aim to demonstrate the superior performance of TS-Reasoner across diverse and challenging tasks that require complex reasoning capabilities.

We primarily benchmarked TS-Reasoner against two advanced approaches. The first is o1-preview, an advanced reasoning model developed by OpenAI. o1-preview is specifically designed for tackling tasks requiring multi-step reasoning and decision-making, leveraging large-scale pretraining and structured reasoning pathways to achieve high accuracy. It has demonstrated significant success in tasks requiring complex problem decomposition and logical reasoning, making it an ideal baseline for our evaluation.

The second approach we considered is based on the ReAct framework. This reasoning structure takes inspiration from the dynamic interplay between "reasoning" and "acting," mimicking human behavior when acquiring new skills and solving problems. By integrating reasoning directly into the action process, ReAct is capable of handling tasks requiring adaptive learning and efficient decision-making, which has made it a popular framework for reasoning-based AI systems.

As shown in Table 4,5,10, TS-Reasoner outperforms both baselines in decision-making tasks, compositional QA tasks, as well as causal mining tasks. The experimental result further validated TS-Reasoner as a simple but effective solution to multi-step reasoning in domain specific time series practical application scenarios.

Task Requirement	**TS-Reasoner**			**o1-preview**			**ReAct**		
	SR(%)	**AAP**	**RAP**	**SR(%)**	**AAP**	**RAP**	**SR(%)**	**AAP**	**RAP**
Profit Percent	**59.2**	**243.31**	**32.34**	6.1	12.53	-198.43	0.0	-	-
Risk Tolerance	**96.0**	54.54	-46.04	18.0	**124.72**	**24.14**	4.0	-0.04	-100.63
Budget Allocation	**90.0**	**37.12**	**7.57**	28.0	-195.96	-225.50	4.0	15.70	-13.84

Table 4: The success rate and performance of TS-Reasoner against additional baselines on desicion making. SR stands for Success Rate; AAP stands for Absolute Average Profit. RAP is the Relative Average Profit compared to vanilla strategy. In Profit Percent and Budget Allocation task, we aim at improving the profit. Thus positive RAP is expected. In Risk Tolerance, the model is required to first ensure the risk and minimize the profit reduction. A negative RAP indicates a more conservative model in terms of risk management compared to vanilla strategy. **Bold** indicates the best results.

Task	**Reasoning Steps**	**TS-Reasoner**		**o1-preview**		**ReAct**	
		SR(%)	**MAPE(std)**	**SR(%)**	**MAPE(std)**	**SR(%)**	**MAPE(std)**
Stock Future Price Prediction	1	**100.0**	**0.042(0.030)**	100.0	0.053(0.031)	48.00	0.043(0.023)
Stock Future Volatility Prediction	2	**100.0**	**0.748(0.691)**	100.0	0.750(0.533)	46.00	1.123(0.882)
Energy Power w/ Max Load	3	**97.87**	0.101(0.339)	78.72	**0.095(0.198)**	21.28	0.136(0.292)
Energy Power w/ Min Load	3	**97.83**	**0.084(0.104)**	76.09	0.218(0.352)	36.96	0.374(0.796)
Load Ramp Rate in Energy Power	3	**100.0**	**0.060(0.153)**	91.67	0.076(0.179)	29.17	0.131(0.273)
Load Variability Limit in Energy Power	3	**93.88**	0.288 (0.385)	89.80	**0.169(0.290)**	26.53	0.268(0.360)

Table 5: The overall success rate and performance of our model against additional baselines on compositional QA. SR stands for Success Rate; MAPE is the Mean Absolute Percentage Error. **Bold** indicates the best results.

Task	Reasoning Steps	TS-Reasoner-C		TS-Reasoner-L		TS-Reasoner-L + paraphrased data	
		SR(%)	MAPE(std)	SR(%)	MAPE(std)	SR(%)	MAPE(std)
Stock Future Price Prediction	1	**100.0**	**0.042(0.030)**	**100.0**	**0.042(0.030)**	20.00	0.046(0.030)
Stock Future Volatility Prediction	2	**100.0**	**0.748(0.691)**	**100.0**	**0.748(0.691)**	**100.0**	**0.748(0.691)**
Energy Power w/ Max Load	3	**97.87**	**0.101(0.339)**	**97.87**	**0.101(0.339)**	**97.87**	**0.101(0.339)**
Energy Power w/ Min Load	3	97.83	**0.084(0.104)**	**100.00**	0.086(0.103)	**100.00**	0.086(0.103)
Load Ramp Rate in Energy Power	3	**100.0**	0.060(0.153)	93.75	0.058(0.149)	97.91	**0.053(0.144)**
Load Variability Limit in Energy Power	3	93.88	0.288 (0.385)	**97.96**	**0.203(0.308)**	89.80	0.294(0.375)

Table 6: The overall success rate and performance of TS-Reasoner variants. TS-Reasoner-C denotes TS-Reasoner with ChatGPT as task decomposer leveraging it's in context learning ability, TS-Reasoner-L denotes TS-Reasoner with finetuned LLAMA as task decomposer, TS-Reasoner-L + paraphrased data denotesTS-Reasoner with LLAMA finetuned on para- phrased data as task decomposer evaluated on paraphrased data. SR stands for Success Rate; MAPE is the Mean Absolute Percentage Error. Bold indicates the best results

Task Requirement	TS-Reasoner			o1-preview			ReAct		
	SR(%)	ACC(%)	SSR(%)	SR(%)	ACC(%)	SSR(%)	SR(%)	ACC(%)	SSR(%)
Causal Relationship	**100.0**	**79.15**	**8.0**	82.0	74.08	4.0	50.0	76.13	0.0

Table 7: The success rate and performance of TS-Reasoner against other baselines on causal relationship recognition. SR stands for Success Rate; ACC stands for Accuracy; SSR stands for Strict Success Rate. **Bold** indicates the best results.

Dataset	Number of CSVs	Avg Total Timestamps	Number of Variables
Daily Yahoo Stock	6780	3785	7
Hourly Yahoo Stock	5540	35	7
Energy Data	66	872601	11
Causal Data	8	529	3–6

Table 8: Dataset Statistics of the constructed dataset. The exact number of time series are not calculated because it depends on randomly sampled sequence length when generating task instances.

Task Requirement	w/ Granger			w/ Bayesian			w/ LiNGAM			w/ Causal Forest		
	SR(%)	ACC(%)	SSR(%)	SR(%)	ACC(%)	SSR(%)	SR(%)	ACC(%)	SSR(%)	SR(%)	ACC(%)	SSR(%)
Causal Relationship	**100.0**	**79.15**	8.0	**100.0**	58.61	0.0	**100.0**	62.10	0.0	90.0	74.81	**12.0**

Table 10: The success rate and performance of TS-Reasoner with different causal tools

Task	TS-Reasoner-C		TS-Reasoner-L	
	Avg Input	Avg Output	Avg Input	Avg Output
Stock Profit Percent	2670.0	49.0	142.0	49.0
Stock Risk Tolerance	2668.0	50.6	140.0	50.6
Stock Budget Allocation	2676.4	66.0	148.4	66.0
Easy Stock Future Price	2614.0	49.0	86.0	49.0
Easy Stock Future Volatility	2609.0	41.8	81.0	41.8
Easy Stock Future Trend	2613.0	45.0	85.0	45.0
Electricity Prediction Max Load	2657.6	110.6	129.6	110.6
Electricity Prediction Min Load	2654.0	56.6	126.0	56.6
Electricity Prediction Load Ramp Rate	2656.6	75.4	128.6	75.4
Electricity Prediction Load Variability Limit	2658.6	130.0	2658.6	79.0
Causal Relation	2648.2	74.0	120.2	74.0
Average	2647.76	63.36	119.76	63.36

Table 9: Token Analysis for each question type. In-Context denotes TS-Reasoner with ChatGPt 3.5 turbo as backbone leveraging its in-context learning ability. Finetuned denotes TS-Reasoner with LLAMA 3.1 8b Instruct finetuned on our dataset as backbone. The total number of input tokens is roughly slightly smaller than number of tokens for system prompt (57) + in-context examples (2424)+ question (119.76) + format instruction (69) = 2669.76. Due to the nature of tokenizers, repetitively occurring phrases may be tokenized as a single token which causes the total number of input tokens to be slightly smaller than the sum of its parts being tokenized individually.