

A IMPLEMENTATION DETAILS OF TILDE-Q SUBLOSSES

As we discussed in Sec. 4.2, TILDE-Q consists of three sublosses: $L_{a.shift}$, L_{phase} , and L_{amp} . Our design rationale for selecting these sublosses is described in Sec. 4.1. In this section, we describe the detailed connection between the sublosses and the design rationale (Eqs. 1, 2, and 3).

Amplitude Shifting Given two sets of points with the same length T , $X, X' \in \mathbb{R}^T$, let us define their distance using the signed distance function $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$. Then, for each point x, x' in set X, X' , we can define a point-wise distance set D with g as below:

$$D = [g(x_1, x'_1), \dots, g(x_T, x'_T)] = [d_1, \dots, d_T].$$

When we design $L_{a.shift}$, we have one main question: given an arbitrary X, X' , and g , how do we build a loss function that is invariant to *any* arbitrary gap k . In this work, we have reformulated this task from *ensuring equal gaps between all points* into *making uniform distribution of the gaps* (i.e., $\sum_i p_{d_i} \log p_{d_i}$ on the interval $[1, T]$). Please note that we convert gaps into relative values since an absolute domain requires information for k for each data sample. Without loss of generality, we can say that this problem is equivalent to the problem of entropy maximization. Let us suppose that we convert the distance set D into the probability distribution by *Softmax* function, $p_{d_i} = \text{Softmax}(d_i)$. In this case, we can say that our optimization problem is maximize the entropy as below:

$$\text{maximize } L = \sum_{i=1}^T p_{d_i} \log p_{d_i},$$

which is well-known to have its global optima with $\forall_{i \in [1, T]} p_{d_i} = \frac{1}{T}$. Therefore, we formulate $L_{a.shift}$ as Eq. 4, which satisfies Eq. 1. Please note that the noise robustness of $L_{a.shift}$ relies on that of the signed distance function, g . Since $L_{a.shift}$ requires computation of g and *Softmax*, it takes $O(n)$ time for its computation.

Phase Shifting To discuss phase shifting and periodicity of a time-series, Fourier transform is an inevitable factor. However, in the real-world dataset, a few problems arise: 1) we are unaware of the frequencies and periodicity of the data itself, and 2) a direct use of Fourier coefficients may be biased by noise. During the design phase, we aim to solve these problems with L_{phase} . To extract the main flow of time-series data (i.e., the dominant periodicity or frequencies), we first define the dominant frequencies based on their statistical significance. Let $X \in \mathbb{R}^T$ as an input signal. In the machine learning domain, researchers commonly suppose the input signal follows normal distribution $X \sim \mathcal{N}(0, I)$. Accordingly, its Fourier coefficients on frequency k is:

$$\mathcal{F}(X) = \sum_{n=1}^T x_n e^{-i2\pi kn/T} \sim \mathcal{N}(0, T).$$

After Fourier transform, we define k as a dominant frequency if k is greater than \sqrt{T} , which indicates statistical significance. However, in some cases, we have only a short sample to represent signals or a noisy signal that has no periodicity, which does not yield a statistically significant k . To prevent such cases, in L_{phase} , we guarantee that at least \sqrt{T} number of frequencies are selected as dominant frequencies. L_{phase} requires $O(n \log n)$ time for its computation, which is inherited from complexity of Fast Fourier Transform.

Uniform Amplification Although effective, L_{phase} has two limitations: 1) it is not perfectly phase shifting invariant as it is optimized with Fourier coefficients, and 2) aforementioned two subloss terms still make no consideration for uniform amplification invariance. Inspired by Paparrizos & Gravano (2015), we utilize normalized correlation for the uniform amplification. Specifically, we normalize correlation R as follows:

$$R(\mathbf{X}, \mathbf{Y}) = \frac{\text{Corr}(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{Corr}(\mathbf{X}, \mathbf{X}) \cdot \text{Corr}(\mathbf{Y}, \mathbf{Y})}},$$

where *Corr* is cross-correlation or auto-correlation, and R is normalized correlation. By using this term, we have uniform amplification invariant measure. We utilize L_{amp} as the subcomponent with small γ , since tolerance for the multiplication factor (i.e., uniform amplification) has greater influence than addition or phase shifting. As L_{phase} , by using fast Fourier transform, L_{amp} takes $O(n \log n)$ time.

TILDE-Q Design Rationale: α and γ $L_{a.shift}$ is built for amplitude shifting and designed to be effective with both periodic and nonperiodic signals. In contrast, L_{phase} handles uniform amplification and is tailored to perform optimally with periodic signals. Since $L_{a.shift}$ and L_{phase} complement each other, we set α to balance them. For example, a large α value will work well for nonperiodic signals and will have less penalty for amplitude shifting. Additionally, we utilize L_{amp} as a subcomponent to calibrate the results (e.g., $\gamma = 0.01$). With this design, while preserving the shape-awareness of TILDE-Q, users can control specific invariances or conditions. For example, users can increase the value of α to emphasize nonperiodic modeling when a dataset has no particular periodicity. This user-oriented objective setting is one of the strengths of TILDE-Q and increases its utility.

B DETAILED EXPERIMENT SETUP

Dataset In our experiment, we utilize six datasets – Synthetic, ECG5000, and Traffic dataset for the simple model (i.e., Sequence-to-Sequence Gated Recurrent Unit) and ETTh2, ETTm2, and Electricity for the state-of-the-art model (i.e., Informer and N-Beats). For each dataset, we describe some metadata of them and the experimental setting, including the input length n and prediction window L . Our implementation could be found in Anonymized Github¹.

Synthetic: As Le Guen & Thome (2019) describe, the Synthetic dataset is an artificial dataset for measuring model performance on sudden changes (step functions) with an input signal composed of two peaks. The amplitude and temporal position of the two peaks are randomly selected. Then the selected position and amplitude of the step are determined by a peak position and amplitude. We use 500 time-series for training, 500 for validation, and 500 for testing. For the Synthetic dataset, we set the input length as $n = 20$ and the prediction window as $L = 40$. The generation code is provided in DILATE Github².

ECG5000: This dataset is originally a 20-hour long ECG (Electrocardiogram), downloaded from Physionet³ and archived in UCR Time Series Classification Archive (Dau et al., 2019). The data is split by each heartbeat and processed in equal lengths (140). In the training, we use 500 for training, 500 for validation, and 4000 for testing. We take the first $n = 84$ steps as input and predict the last $L = 56$ steps.

Traffic: Traffic dataset is a collection of 48 months (2015-2016) hourly road occupancy rate (between 0 to 1) data from the California Department of Transportation⁴. As Le Guen & Thome (2019) do, we utilize univariate series of the first sensor, a total of 17544 data points. We set our problem as forecasting $L = 24$ future occupancy rates with $n = 168$ historical data (past week). We use 60% of the data for training, 20% for validation, and the rest for evaluation.

ETT: The ETT (Electricity TraNSformer Temperature) dataset, published by Zhou et al. (2021), is 2-year data collected from two separate counties in China, including ETTh2 and ETTm2 datasets. Each data point has a target value of “oil temperature” and other 6 power load features. ETTh2 and ETTm2 datasets have 1-hour and 15-minute intervals, respectively. As Zhou et al. (2021) do, we split them into 12/4/4 months for the training/validation/testing. Detailed settings, such as the input and output length and hyperparameter setting, are based on the information at Informer Github⁵.

ECL: The ECL (Electricity Consuming Load) is a dataset recorded in kWh every 15 minutes from 2012 to 2014, for 321 clients. In our experiment, we split them into 15/3/4 months for the train/validation/test, as Zhou et al. (2021) do. Note that we use the same hyperparameter settings in the ETTh2 dataset.

Deep Learning Model Architectures We perform experiments with three different model architectures, including Sequence-to-Sequence GRU, Informer, and N-Beats. To induce models to predict future time-series in a timely manner, we set $\alpha = 0.5$ and $\gamma = 0.01$ for TILDE-Q. Other training

¹<https://anonymous.4open.science/r/TILDE-Q-9E54>

²<https://github.com/vincent-leguen/DILATE>

³<https://physionet.org/>

⁴<http://pems.dot.ca.gov>

⁵<https://github.com/zhouhaoyi/Informer2020>

metrics, including MSE and DILATE, are used as described in their original papers. All models are trained with Early Stopping and ADAM optimizer.

Sequence-to-Sequence GRU To evaluate TILDE-Q in simple model, we utilize one layer Sequence-to-Sequence GRU model. For the training of the GRU model, we set a learning rate of $1e-3$, hidden size of 128, trained by maximum 1000 epochs with Early Stopping and ADAM optimizer.

Informer When we train Informer with ETTh2, ETTm2, and ECL dataset, we utilize the official code and hyperparameter setting. In the case of ECL dataset, as the author answered in their official code⁵, we utilize the same hyperparameter and dataset splitting criteria as the ETTh2 dataset.

N-Beats For N-Beats, we utilize two generic blocks with a hidden size of 128. Additionally, we set the learning rate as $1e-3$ for all three datasets.

Autoformer For Autoformer⁶, we use the official code and hyperparameter setting. For the ETTh2 dataset, we utilize hyperparameter settings described in the official code of FEDformer⁷.

FEDformer For FEDformer⁷, we use the official code and hyperparameter setting.

C ADDITIONAL EVALUATIONS

C.1 DETAILED EXPERIMENT RESULTS AND ANALYSIS

Table 3: Detailed experimental results on six real-world datasets (four cases) with N-Beats.

Methods		N-Beats + MSE				N-Beats + DILATE				N-Beats + TILDE-Q			
Metric		MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS
ETTh2	96	0.1869	7.2379	2.3787	0.4688	0.3105	6.5849	3.6490	0.4879	0.1557	5.1011	1.3240	0.5862
	192	0.2385	11.5667	4.9153	0.4505	0.6186	9.7254	7.0831	0.4637	0.1738	7.6334	2.4122	0.5819
	336	0.2889	16.5255	11.5207	0.4544	1.1406	13.7328	14.6986	0.4584	0.2132	11.3351	5.3556	0.5373
	720	0.3881	24.1570	18.8462	0.4381	1.6713	19.4392	23.7028	0.4575	0.3044	17.6006	9.6636	0.5287
ETTM2	96	0.0790	3.9685	2.0436	0.6721	0.1524	7.9302	5.5597	0.4379	0.0952	4.0110	2.1939	0.6902
	192	0.1224	6.8695	3.2834	0.5762	0.2055	10.0393	8.5602	0.5107	0.1286	6.3556	4.9798	0.6160
	336	0.1824	12.1438	8.5915	0.4587	0.2501	12.6342	16.1473	0.4819	0.1705	8.9377	8.3539	0.6195
	720	0.2370	22.8676	17.8458	0.4929	0.4170	17.7764	24.6877	0.5836	0.2336	14.2715	19.0883	0.7070
ECL	96	0.3666	3.5207	0.2989	0.6589	1.1156	5.1430	2.6613	0.5074	0.3183	2.9707	0.4844	0.7229
	192	0.4307	5.7578	0.4253	0.6212	1.1859	7.3406	2.8488	0.4973	0.3383	4.1817	0.4229	0.7187
	336	0.5199	8.5563	0.5384	0.5965	1.2460	9.5096	3.0517	0.5091	0.3831	5.6643	0.3024	0.7112
	720	0.6240	13.9436	0.6510	0.5717	1.3061	13.1928	3.7279	0.5337	0.4540	8.9997	0.3251	0.6960
Exchange	96	0.4496	8.6395	4.3197	0.4424	0.3945	8.9661	4.3286	0.4316	0.2748	7.9744	5.2964	0.4467
	192	1.2161	12.1857	10.5166	0.4157	1.5684	13.0560	9.2434	0.4061	1.6629	11.5557	8.7896	0.4348
	336	1.4529	14.7085	19.0407	0.4130	3.6784	17.5189	17.2512	0.3871	1.8432	12.5648	20.8871	0.4603
	720	1.8563	21.7347	50.6751	0.4073	3.9008	26.7020	74.0546	0.3400	2.8487	19.1588	53.8069	0.4619
Traffic	96	0.2349	2.1046	0.0216	0.8303	2.3325	3.9657	1.2052	0.5250	0.2286	2.0699	0.0207	0.8371
	192	0.3014	3.4040	0.0142	0.7916	2.5627	5.4169	1.1355	0.5515	0.3352	3.2559	0.0119	0.8028
	336	0.3455	4.6409	0.0088	0.7918	2.4599	8.2828	1.3377	0.5208	0.3990	4.2622	0.0066	0.8206
	720	0.4298	7.0561	0.0045	0.7958	2.3522	12.6258	0.9967	0.5177	0.4480	6.7344	0.0034	0.8085
Weather	96	0.0042	9.3228	5.9134	0.4072	0.0023	8.9289	5.0617	0.4256	0.0010	6.5198	6.0450	0.5168
	192	0.0056	10.9682	11.9549	0.4212	0.0030	12.8164	10.6858	0.4307	0.0017	8.8391	9.0867	0.5076
	336	0.0058	13.3578	14.6572	0.4243	0.0087	20.4895	23.7903	0.3579	0.0026	11.8682	9.6758	0.5074
	720	0.0068	18.5861	22.1432	0.4315	0.1534	28.6021	47.4488	0.3982	0.0029	17.1895	19.1942	0.5078

At first, we observe that the model optimized with TILDE-Q outperforms the same model optimized with other objective functions in both short- and long-term forecasting tasks. An interesting point in the results is the large increased errors of TDI and DTW with long-term forecasting. For example, TDI of Informer with DILATE shows dramatically increased error with the ECL dataset, as the forecasting window increases, while LCSS does not produce such a large increased error. We attribute this to the weakness of DTW-based loss functions, which have a weakness due to high sensitiveness to noise. In contrast, TILDE-Q does not show such a large performance drop and even achieves

⁶<https://github.com/thuml/Autoformer>

⁷<https://github.com/MAZiqing/FEDformer>

Table 4: Detailed experimental results on six real-world datasets (four cases) with Informer.

Methods		Informer + MSE				Informer + DILATE				Informer + TILDE-Q			
Metric		MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS
ETTh2	96	0.2466	6.9254	3.6676	0.4633	0.3284	6.3109	3.5838	0.5037	0.1768	5.8437	1.6734	0.5379
	192	0.2818	10.2654	11.1580	0.4254	0.4086	8.8262	7.1780	0.4893	0.2432	10.2134	9.9865	0.4317
	336	0.3089	12.1822	18.7014	0.4434	0.4164	10.3779	13.2580	0.5062	0.2958	13.5586	20.2850	0.4165
	720	0.2877	17.6369	38.4617	0.4425	0.4229	14.1196	23.9403	0.4815	0.3157	18.4617	43.3238	0.4262
ETTm2	96	0.0889	3.4007	1.5719	0.7386	0.1263	6.0144	2.7757	0.5129	0.0871	3.1354	1.3474	0.7817
	192	0.1157	5.7964	2.8128	0.6705	0.2340	9.7004	7.8354	0.5266	0.1317	5.7093	2.9129	0.6983
	336	0.1860	8.9971	6.7970	0.6365	0.2805	11.7889	13.3861	0.5025	0.1767	9.0866	7.4023	0.6555
	720	0.2165	14.7685	24.6694	0.5768	0.3745	16.7734	29.2783	0.4747	0.2063	15.3057	24.1959	0.5860
ECL	96	0.2709	2.8067	0.1720	0.7032	0.9856	3.6394	1.4794	0.6324	0.2800	2.9466	0.2473	0.7275
	192	0.2793	4.1193	0.1508	0.7060	1.1209	5.2289	2.1749	0.6053	0.3077	4.2693	0.2978	0.7336
	336	0.3203	5.9533	0.1642	0.7222	1.2331	7.8470	3.0415	0.5694	0.3271	5.8090	0.1984	0.7143
	720	0.6414	15.8561	4.4284	0.4564	1.3706	12.5981	5.6720	0.5506	0.4676	11.4027	0.7107	0.6298
Exchange	96	0.3534	8.0965	4.8843	0.4689	0.3260	7.7370	5.6336	0.4678	0.5264	7.9866	6.5120	0.4553
	192	0.9682	11.0843	11.3110	0.4647	0.9737	10.8894	15.6770	0.4584	1.2845	10.4358	10.7009	0.4959
	336	1.3710	12.8076	18.5937	0.4676	1.6735	12.7034	29.2013	0.4428	1.6912	12.2349	18.2197	0.4932
	720	1.7586	22.6852	59.4243	0.4681	1.8292	16.0093	56.8687	0.5293	1.9130	24.0510	62.8152	0.5104
Traffic	96	0.2606	2.0994	0.0208	0.8329	2.9612	2.3355	0.9646	0.7312	0.2284	2.0027	0.0194	0.8490
	192	0.2920	3.2573	0.0126	0.8158	2.9978	3.5451	0.8429	0.7394	0.2753	3.1721	0.0125	0.8248
	336	0.3109	4.6581	0.0078	0.8115	2.9696	4.9879	1.2672	0.7117	0.2993	4.4715	0.0077	0.8170
	720	0.3472	6.7989	0.0040	0.8146	2.6845	10.7450	3.4514	0.5874	0.3859	7.5424	0.0051	0.7752
Weather	96	0.0043	8.2890	5.4604	0.4556	0.0069	6.5571	4.7505	0.5159	0.0021	5.5412	4.5012	0.5602
	192	0.0031	10.7993	9.2928	0.4523	0.0041	10.5645	9.4713	0.4704	0.0028	8.2535	5.8289	0.5516
	336	0.0051	13.8721	22.2699	0.4451	0.0055	12.0586	16.4933	0.4884	0.0039	10.8802	10.5220	0.5668
	720	0.0061	21.7720	41.5877	0.4476	0.0737	16.8378	29.8112	0.5142	0.0047	13.9934	20.9991	0.5689

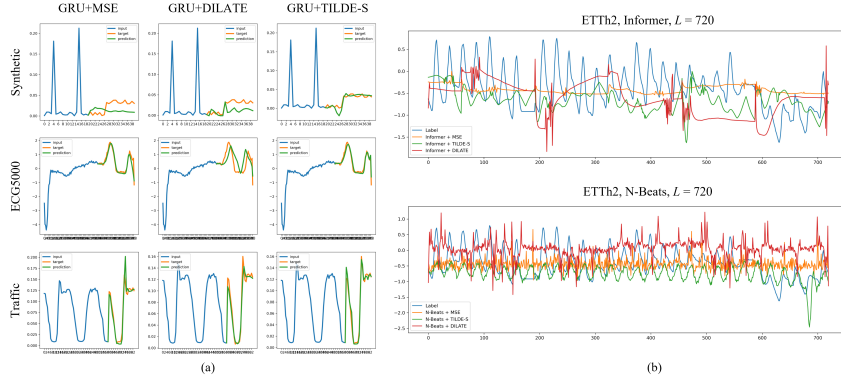


Figure 3: Qualitative results with simple sequence-to-sequence GRU model (a) and state-of-the-art model (b).

better performance in the long-term forecasting (e.g., Table 4, ETTh2). Additionally, we can find that Informer with TILDE-Q on ECL data and N-Beats with TILDE-Q on all three datasets show significant improvements. It indicates that TILDE-Q success to model *shape*, but other metrics could not. We provide additional qualitative analysis in Appendix C.2. **For the experimental results with standard deviation and NLinear (Zeng et al., 2023), please visit our Anonymous GitHub¹.**

Next, we present a qualitative analysis of the results. Fig. 3 shows how the model with different training metrics forecasts with different datasets. From the figure, we have noticed that TILDE-Q allows the model to generate more robust, shape-aware forecasting, regardless of amplitude shifting, phase shifting, and uniform amplification. For example, in the case of N-Beats (Fig. 3 (b) bottom), TILDE-Q generate forecasting results, which are more robust, shape-aware prediction compared to other metrics. We also see the strength in the Informer case (Fig. 3 (b), top). Even when the model has not enough ability to capture shape, TILDE-Q tries to retrieve the shape. We provide additional qualitative results with the visualization below. When the model has enough ability to

capture shape (i.e., except ETTh2, Informer of $T' \in [192, 336, 720]$), TILDE-Q has shown its noise-robust, smooth forecasting with correctly modeled temporal dynamics. In the most of N-Beats results and some of the Informer results, TILDE-Q reveals that these models have enough ability to capture the temporal dynamics with a proper loss function. In summary, TILDE-Q proves that it is model-agnostic, noise-robust, and able to capture the shape.

Table 5: Detailed experimental results on six real-world datasets (four cases) with Autoformer.

Methods		Autoformer + MSE				Autoformer + DILATE				Autoformer + TILDE-Q			
Metric		MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS
ETTh2	96	0.1538	5.2227	2.1865	0.6187	0.2211	6.0453	2.5345	0.5315	0.1494	5.1060	1.9752	0.6317
	192	0.1974	7.8730	3.3382	0.6019	0.2825	8.6696	5.6671	0.5335	0.2079	7.8917	3.7532	0.5984
	336	0.2393	10.8002	7.3141	0.5954	0.3759	11.0335	13.1347	0.5257	0.2360	10.7212	7.0085	0.5971
	720	0.2859	16.3502	15.9233	0.5772	0.4296	15.9819	22.2173	0.4924	0.2378	16.0002	13.7906	0.5795
ETTm2	96	0.0990	4.3498	2.5052	0.6756	0.1135	5.3097	2.2211	0.5936	0.0940	3.9078	2.2587	0.7075
	192	0.1340	6.3207	3.3676	0.6512	0.1854	8.5209	3.7894	0.5506	0.1259	6.0979	2.9278	0.6810
	336	0.1587	9.4374	6.9205	0.6036	0.2001	12.0265	8.8305	0.5370	0.1548	9.5223	7.2875	0.6169
	720	0.1999	14.8332	11.9655	0.6064	0.2665	17.8025	17.4114	0.5001	0.1885	14.5844	9.9918	0.6277
ECL	96	0.4209	3.5957	0.2461	0.6487	0.6813	3.6490	0.4780	0.6253	0.3515	3.2173	0.2298	0.6912
	192	0.4206	4.9924	0.3416	0.6574	0.7319	5.5324	0.2775	0.6118	0.4032	4.8581	0.3301	0.6680
	336	0.4621	6.6888	0.2795	0.6535	0.7895	7.5665	0.2503	0.6091	0.4637	6.7335	0.3923	0.6429
	720	0.5005	10.8571	0.2383	0.6183	0.8630	12.1416	0.1877	0.6074	0.5049	9.8492	0.2525	0.6420
Exchange	96	0.2472	8.2957	5.8340	0.4577	0.1921	8.4651	5.6328	0.4646	0.1730	8.2046	5.1165	0.4577
	192	0.3255	11.4212	17.0909	0.4319	0.4732	12.8599	19.0164	0.4124	0.2955	11.3655	15.4372	0.4433
	336	0.5483	15.1853	44.4975	0.3277	0.8035	18.0948	57.5819	0.3114	0.5331	16.7350	45.8166	0.3321
	720	1.3620	24.6397	145.3080	0.2357	1.4936	27.7069	151.6671	0.2302	1.1993	19.5296	121.8509	0.2233
Traffic	96	0.2562	1.9689	0.0178	0.8761	0.4835	1.9044	0.0392	0.8521	0.2275	1.8778	0.0168	0.8879
	192	0.2604	2.8922	0.0091	0.8780	0.5653	3.0466	0.0343	0.8187	0.2497	2.8793	0.0116	0.8817
	336	0.2474	4.0026	0.0051	0.8797	0.8155	4.2637	0.0327	0.8047	0.2422	3.9469	0.0059	0.8760
	720	0.2720	6.4371	0.0030	0.8710	1.0729	6.0776	0.0217	0.8176	0.2836	6.1751	0.0034	0.8674
Weather	96	0.0168	7.4658	6.0336	0.4818	0.0019	5.9775	4.9688	0.5306	0.0015	5.9829	4.8957	0.5461
	192	0.0069	10.6173	7.4506	0.4941	0.0026	8.3686	5.4565	0.5423	0.0017	7.7799	6.1032	0.5355
	336	0.0052	12.5224	13.2607	0.4898	0.0030	12.4524	12.1816	0.4854	0.0020	10.3144	9.0025	0.5252
	720	0.0078	18.5079	25.8063	0.4744	0.0115	20.1354	36.7754	0.4721	0.0023	15.2563	18.3134	0.5102

Table 6: Detailed experimental results on six real-world datasets (four cases) with FEDformer.

Methods		FEDformer + MSE				FEDformer + DILATE				FEDformer + TILDE-Q			
Metric		MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS
ETTh2	96	0.1299	4.7265	1.2607	0.6690	0.1906	6.2294	1.8228	0.5261	0.1381	4.7578	1.3560	0.6621
	192	0.1819	7.6178	2.6979	0.6229	0.2688	8.8422	4.8043	0.5261	0.1988	7.6174	2.7712	0.6124
	336	0.2305	10.5860	6.7027	0.6050	0.3506	11.4834	12.8408	0.5091	0.2382	10.4108	6.6218	0.6039
	720	0.2776	15.7013	14.7466	0.5911	0.4327	14.0692	20.6266	0.5091	0.2871	15.3120	16.4059	0.5808
ETTm2	96	0.0682	3.0962	1.3862	0.7868	0.1147	4.6648	2.2981	0.6325	0.0669	3.0328	1.3556	0.7918
	192	0.0976	5.2417	2.0295	0.7340	0.1848	8.0678	4.4893	0.5391	0.0971	5.1508	2.1782	0.7384
	336	0.1326	8.3151	5.4619	0.6667	0.2493	13.6349	11.7563	0.5049	0.1279	8.3010	4.5488	0.6828
	720	0.1957	14.2579	11.8328	0.6262	0.2913	17.4636	41.9434	0.4806	0.1822	14.1131	10.4778	0.6361
ECL	96	0.2531	2.6402	0.1436	0.7322	0.4794	2.8685	0.2482	0.6943	0.2638	2.6594	0.1614	0.7265
	192	0.2945	3.8647	0.1831	0.7306	0.5485	4.3313	0.1732	0.6813	0.2821	3.7830	0.1277	0.7340
	336	0.3313	5.2789	0.1078	0.7207	0.6967	5.7911	0.1985	0.6892	0.3385	5.1763	0.1229	0.7290
	720	0.3956	8.5881	0.0632	0.6961	0.7741	10.1163	0.8837	0.6403	0.3939	8.5665	0.0784	0.7013
Exchange	96	0.1437	8.5595	4.4258	0.4347	0.3884	8.9178	7.0385	0.4439	0.1215	8.0591	6.1979	0.4704
	192	0.2694	12.5168	12.3117	0.4202	0.5912	13.1929	15.4207	0.4187	0.2956	11.5607	12.1302	0.4474
	336	0.4916	16.4673	27.0756	0.4140	0.7520	18.1381	33.7878	0.3969	0.5896	15.9267	24.8808	0.4342
	720	1.2115	25.6243	108.0500	0.3838	1.5110	26.7031	91.6313	0.3760	1.1700	24.6190	71.8783	0.3935
Traffic	96	0.2074	1.9132	0.0165	0.8824	0.3533	1.8617	0.0291	0.8609	0.1867	1.8386	0.0152	0.8983
	192	0.2051	2.7761	0.0085	0.8951	1.4682	2.7545	0.1312	0.8591	0.1961	2.7380	0.0083	0.8975
	336	0.2140	3.7583	0.0047	0.9023	2.9741	3.7440	0.1779	0.8519	0.2059	3.7834	0.0050	0.8947
	720	0.2291	5.9735	0.0026	0.8919	3.0829	5.8173	0.0949	0.8580	0.2312	6.1080	0.0022	0.8735
Weather	96	0.0070	6.1550	4.7039	0.5264	0.0017	5.9507	4.5891	0.5535	0.0014	5.5205	4.1085	0.5792
	192	0.0063	7.6906	6.2940	0.5417	0.0020	8.6593	4.6333	0.6000	0.0017	6.9706	5.0003	0.5863
	336	0.0046	10.3501	9.4691	0.5261	0.0053	14.8249	9.6239	0.4801	0.0018	9.3016	7.4610	0.5784
	720	0.0060	15.2429	24.5379	0.4911	0.0029	16.3929	20.0782	0.4892	0.0024	13.7936	15.7668	0.5737

Table 7: Detailed experimental results on six real-world datasets (four cases) with NSFormer.

Methods		NSFormer + MSE				NSFormer + TILDE-Q			
Metric		MSE	DTW	TDI	LCSS	MSE	DTW	TDI	LCSS
ETTh2	96	0.1868 \pm 0.0173	5.7546 \pm 0.5472	2.0024 \pm 0.1875	0.5522 \pm 0.0504	0.1629 \pm 0.0083	5.1673 \pm 0.0355	1.4536 \pm 0.5094	0.6083 \pm 0.0080
	192	0.2311 \pm 0.0053	8.6408 \pm 0.6911	3.8789 \pm 0.0671	0.5301 \pm 0.0440	0.2016 \pm 0.0056	8.0815 \pm 0.0596	4.0217 \pm 0.1380	0.5963 \pm 0.0004
	336	0.2624 \pm 0.0167	11.1345 \pm 0.3661	10.0050 \pm 0.7225	0.5530 \pm 0.0131	0.2760 \pm 0.0136	11.0140 \pm 0.0735	10.8131 \pm 0.7981	0.5598 \pm 0.0050
	720	0.2391 \pm 0.0025	15.9557 \pm 0.0797	18.2158 \pm 1.4161	0.5409 \pm 0.0154	0.2663 \pm 0.0102	15.0010 \pm 0.1701	19.5509 \pm 1.3960	0.5575 \pm 0.0084
ETTm2	96	0.0682 \pm 0.0011	3.2222 \pm 0.0710	1.5062 \pm 0.0636	0.7681 \pm 0.0029	0.0704 \pm 0.0017	3.2704 \pm 0.0382	1.5464 \pm 0.0593	0.7678 \pm 0.0018
	192	0.1107 \pm 0.0061	5.6567 \pm 0.1432	3.1250 \pm 0.3309	0.6897 \pm 0.0052	0.1142 \pm 0.0097	5.4756 \pm 0.0160	3.0431 \pm 0.1125	0.7091 \pm 0.0051
	336	0.1655 \pm 0.0156	9.0536 \pm 0.2199	7.3496 \pm 0.2607	0.6285 \pm 0.0146	0.1590 \pm 0.0033	8.9267 \pm 0.1396	8.5419 \pm 0.8318	0.6450 \pm 0.0049
	720	0.2349 \pm 0.0202	14.8967 \pm 0.3186	19.1985 \pm 1.7474	0.5587 \pm 0.0156	0.2209 \pm 0.0126	14.8251 \pm 0.0936	20.4820 \pm 0.6624	0.5892 \pm 0.0030
ECL	96	0.3117 \pm 0.0102	3.1460 \pm 0.0466	0.3059 \pm 0.0078	0.6839 \pm 0.0051	0.3136 \pm 0.0069	3.0445 \pm 0.0511	0.3170 \pm 0.0273	0.6876 \pm 0.0012
	192	0.3789 \pm 0.0160	4.6092 \pm 0.0077	0.3790 \pm 0.0398	0.6736 \pm 0.0252	0.3453 \pm 0.0158	4.4510 \pm 0.1726	0.3892 \pm 0.0469	0.6853 \pm 0.0049
	336	0.3856 \pm 0.0072	6.2827 \pm 0.0620	0.2954 \pm 0.0678	0.6724 \pm 0.0098	0.3633 \pm 0.0062	6.0038 \pm 0.0717	0.3573 \pm 0.0336	0.6953 \pm 0.0075
	720	0.4102 \pm 0.0426	10.0263 \pm 0.0869	1.7172 \pm 1.0628	0.6680 \pm 0.0107	0.3999 \pm 0.0071	9.6903 \pm 0.0497	0.6640 \pm 0.2962	0.6716 \pm 0.0059
Exchange	96	0.1514 \pm 0.0505	8.4840 \pm 0.1408	6.2843 \pm 0.3800	0.4600 \pm 0.0100	0.1078 \pm 0.0084	8.0182 \pm 0.2299	7.4383 \pm 0.6279	0.4667 \pm 0.0037
	192	0.2514 \pm 0.0280	10.7493 \pm 0.8153	14.7287 \pm 2.0162	0.4723 \pm 0.0105	0.2219 \pm 0.0247	11.0704 \pm 0.5110	15.5720 \pm 3.1424	0.4898 \pm 0.0191
	336	0.5246 \pm 0.0908	12.3626 \pm 1.0148	25.7294 \pm 1.0591	0.4873 \pm 0.0138	0.4219 \pm 0.0740	11.9023 \pm 0.5967	23.2522 \pm 1.5396	0.4928 \pm 0.0049
	720	0.8866 \pm 0.1059	18.8469 \pm 3.5656	54.3978 \pm 10.9369	0.5006 \pm 0.0075	0.7557 \pm 0.0535	13.7467 \pm 1.3978	42.9306 \pm 9.0610	0.5199 \pm 0.0201
Traffic	96	0.1916 \pm 0.0131	1.7903 \pm 0.0393	0.0148 \pm 0.0006	0.9012 \pm 0.0076	0.2035 \pm 0.0101	1.7658 \pm 0.0213	0.0134 \pm 0.0007	0.9088 \pm 0.0069
	192	0.1829 \pm 0.0086	2.7411 \pm 0.0295	0.0083 \pm 0.0001	0.9000 \pm 0.0040	0.1838 \pm 0.0113	2.7329 \pm 0.0139	0.0079 \pm 0.0002	0.9038 \pm 0.0133
	336	0.2017 \pm 0.0046	3.7920 \pm 0.0107	0.0049 \pm 0.0002	0.8959 \pm 0.0021	0.1919 \pm 0.0054	3.7476 \pm 0.0310	0.0046 \pm 0.0001	0.8993 \pm 0.0019
	720	0.2294 \pm 0.0077	6.0884 \pm 0.1178	0.0027 \pm 0.0001	0.8833 \pm 0.0077	0.2287 \pm 0.0071	5.8202 \pm 0.0803	0.0024 \pm 0.0001	0.8972 \pm 0.0030
Weather	96	0.0017 \pm 0.0005	8.2914 \pm 1.1221	7.8379 \pm 1.5827	0.4230 \pm 0.0378	0.0012 \pm 0.0001	5.7786 \pm 0.0468	5.4118 \pm 0.1960	0.5451 \pm 0.0037
	192	0.0020 \pm 0.0002	9.9842 \pm 0.4992	10.8039 \pm 1.9632	0.4536 \pm 0.0025	0.0014 \pm 0.0000	7.3667 \pm 0.0967	6.2922 \pm 0.2090	0.5532 \pm 0.0060
	336	0.0018 \pm 0.0002	13.6885 \pm 0.7604	14.9242 \pm 1.3217	0.4308 \pm 0.0161	0.0015 \pm 0.0001	9.9883 \pm 0.0800	9.0103 \pm 0.5340	0.5529 \pm 0.0020
	720	0.0022 \pm 0.0002	16.7372 \pm 0.5878	33.7388 \pm 4.4616	0.4751 \pm 0.0025	0.0020 \pm 0.0000	14.7109 \pm 0.4148	17.1988 \pm 0.9428	0.5405 \pm 0.0051

C.2 QUALITATIVE EVALUATION

As we have described, *shape* is hard to be captured by using existing metrics, such as L_p metrics. We therefore showcase each model, training metric, and dataset. For the evaluation of TILDE-Q, we visualize the forecasting results using various models. Specifically, we report four cases: 1) when TILDE-Q and MSE show marginal performance, 2) case to prove that TILDE-Q is scalable for the complex problem, 3) case to visually investigate superiority of TILDE-Q for periodic and noisy dataset, and 4) when the data have no obvious periodicity. For more visualization, and the experimental results, please refer to the Anonymous GitHub¹.

C.2.1 COMPARISON WITH MSE AND TILDE-Q

In some cases, including FEDformer model with ETTh2 dataset, TILDE-Q and MSE have marginal MSE performance, as shown in Fig. 4. In these cases, MSE and TILDE-Q have shown almost identical behavior, indicating that TILDE-Q’s performance is at least comparable to that of MSE.

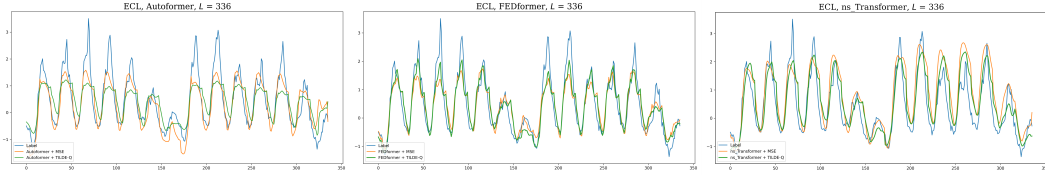


Figure 4: Qualitative results with ECL dataset for Autoformer (left), FEDformer (middle), and NSformer (right). Green line is trained with TILDE-Q and blue line is trained with MSE

C.2.2 USING TILDE-Q TO BOOST MODEL SCALABILITY

In the cases of NBeats, Informer, and even NSformer (Liu et al., 2022), we have found that TILDE-Q is the best methods for model to have better interpretation of dataset, as shown in Fig. 5. These showcases reveal the importance of shape-awareness, especially for a dataset that is hard to predict. For NBeats and Informer, especially for the long-term forecasting (e.g., 720-Output), TILDE-Q shows its superiority in terms of shape-awareness and noise-robustness. For NSformer, TILDE-Q captures peaks and plateaus better than MSE, which is crucial for real-world problem, such as electricity usage. In case of Informer (middle), MSE and TILDE-Q make informative predictions on the 336-Output

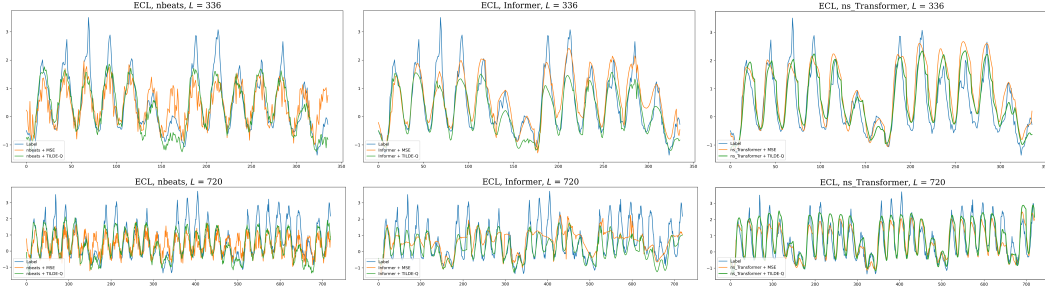


Figure 5: Qualitative results with NBeats (left), Informer (middle), and NSformer (right)

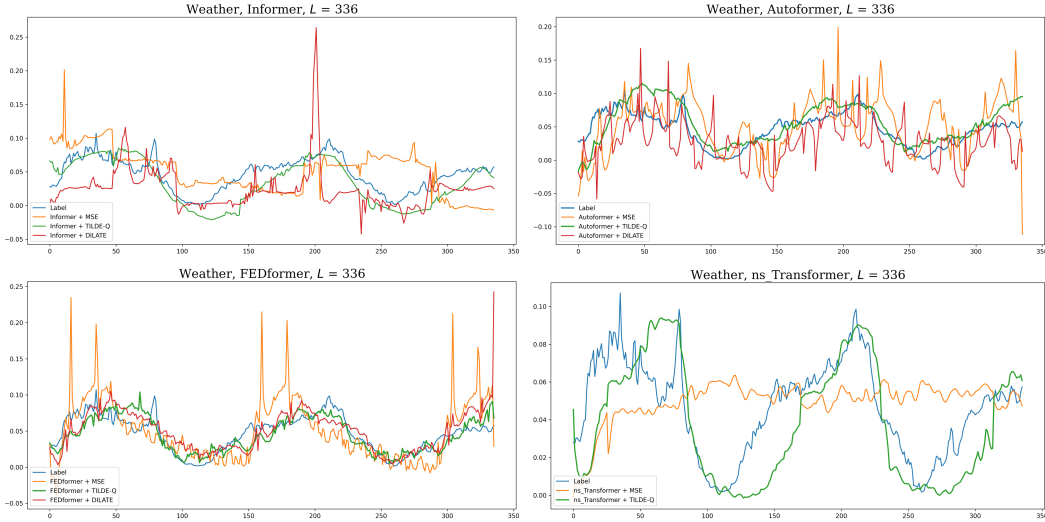


Figure 6: Qualitative results with Weather dataset for Informer (top-left), Autoformer (top-right), FEDformer (bottom-left), and NSformer (bottom-right)

setting, but on the 720-Output setting, the model trained with MSE lost its ability to be aware of the shape, resulting in uninformative output.

C.2.3 PERIODIC AND NOISY DATASET SHOWCASES

Because of its noisy characteristics, most models have struggled to recognize the periodicity of the Weather dataset, resulting in performance degradation. This section will provide qualitative examples from the Weather dataset. In the Weather dataset, we observed that it has periodic behavior with large fluctuations and high noises. This fluctuation makes the model mislead the periodicity of dataset, even when we could visually identify it. As shown in Fig. 6, compared to MSE and DILATE, TILDE-Q helps the model properly forecast the future value, preserving important features such as plateau, peak, and periodicity with less noise. On the other hand, in the cases of MSE and DILATE, they have less effective or no strategy for handling the shape and distortion, yielding uninformative and noisy forecasting results.

C.2.4 PERFORMANCE ON DATA WITHOUT PERIODICITY

Although effective, TILDE-Q and the existing metrics still face limitations when predicting data without obvious periodicity and with rapid fluctuations. One representative example is the Exchange dataset, which inherently has no obvious periodicity, like other economics datasets. However, the results with TILDE-Q and NSformer provide insight for a possible future improvement through its temporal feature modeling with a large gap (i.e., amplitude shifting), as bottom-right part of Fig. 7.

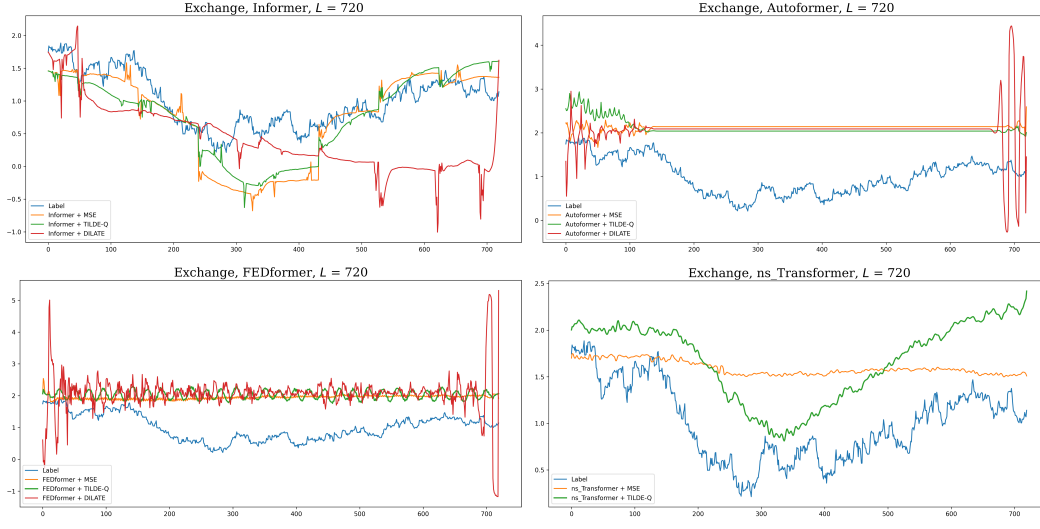


Figure 7: Qualitative results with Exchange dataset for Informer (top-left), Autoformer (top-right), FEDformer (bottom-left), and NSformer (bottom-right)

C.3 ABLATION STUDY

To evaluate the effect of the α , γ , and measure the effect of each loss function, we conduct a set of experiments with the ETTh2 dataset and N-Beats on the long-term forecasting problem. As we can see in Fig. 8, the model tends to predict amplification-free forecasting when α increases. These results indicate our motivation, “ $\mathcal{L}_{a.shift}$ will return the forecasting results with same standard deviation with timely manner but without consideration of proper average value.”

Furthermore, in the top of Fig. 9, we can observe three things: (1) if we utilize $\mathcal{L}_{a.shift}$ only, as we intended, it has a different average (-1.19 vs. 0.11) but relatively similar standard deviation (0.408 vs. 0.299); (2) In the case of \mathcal{L}_{phase} only, they can capture dominant frequency and produce relatively less-noisy forecasting; (3) \mathcal{L}_{amp} have relatively similar average value (-1.195 vs. -0.319), but it has far different standard deviation (0.408 vs. 8.592). In contrast, forecasting results of the model trained with MSE is very noisy and hard to interpret (Fig. 9, bottom). Note that we normalized the results in Fig. 9 because of the scale issue.

In Table 8, we provide how model performances vary with respect to hyperparameters of TILDE-Q. For the default setting, we utilized $\alpha = 0.5, \gamma = 0.01$. Because the design of TILDE-Q mainly focuses on shape modeling, we can see that DTW and LCSS are not critically changing for the hyperparameter. But their trade-offs are revealed in the MSE and TDI. For example, when we decrease α , we can observe TDI increases. It indicates the trade-offs of phase shifting invariance, which has tolerance for non-timely forecasting. Also, we can see that increasing α or γ affects the MSE. When we have $\alpha = 1$, we have no \mathcal{L}_{phase} and less penalty for the statistical differences, and its absence causes the high MSE, as we can see in Fig. 8. γ also affects the MSE, but \mathcal{L}_{phase} reduces \mathcal{L}_{amp} ’s side effect.

Table 8: Ablation study on with ETTh2, $L = 720$, and N-Beats

Metric	Default	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1.0$	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.8$	$\alpha = 1.0$	$\mathcal{L}_{a.shift}$ only	\mathcal{L}_{phase} only	\mathcal{L}_{amp} only
MSE	0.3005	0.2968	0.3083	0.3168	0.3075	0.3161	0.2872	1.1752	1.5123	0.3391	1.8453
DTW	17.5154	17.5265	17.5649	17.7302	17.7564	17.5931	17.6508	17.6886	17.7261	17.848	18.0261
TDI	9.2197	9.1303	9.2261	9.4366	10.3550	9.8957	8.4725	8.7118	10.2519	12.8568	10.5602
LCSS	0.5382	0.5366	0.5277	0.5137	0.5050	0.5137	0.5584	0.5445	0.5341	0.4920	0.5086

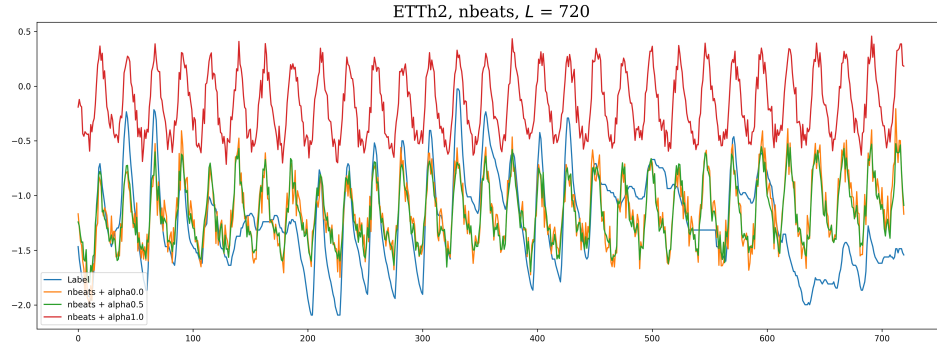
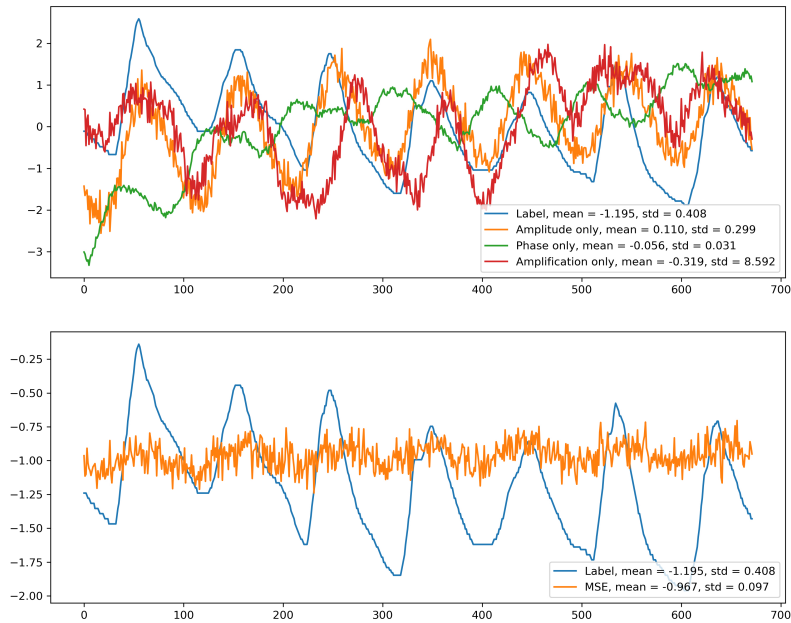
Figure 8: Ablation study result visualization with different α on ETTh2 dataset

Figure 9: Ablation study result visualization of three proposed loss function on ETTh2 dataset