A Training Details

For all models, we trained all models using Standard scaler, Adam optimizer Kingma & Ba (2015) with β_1 = 0.9 and β_2 = 0.999, a batch size of 16, and a constant learning rate of 0.0001 for all settings. We used the length of input sequence L_{in} as 100 for all settings. Also, we conducted experiments varying the length of output sequence L_{out} from 100 to 400. Regarding the coefficients λ_F and λ_R, we set them to 1.0 for all datasets. For λ_E and λ_C in the loss, we used 0.1 for MSL and 1.0 for others. For Synthetic Anomaly Prompting, we adopted the length of anomaly prompts L_z and the size of anomaly prompt pool 5 and 10 respectively as default. In the selection of top-N anomaly prompts in anomaly prompt pool, we used $N = 3$. We trained the models for 5 epochs, with 3 layers of transformer backbone, and the embedding dimension D is fixed to 256. Also, our experiments were executed on single GPU (NVIDIA RTX 3090), implementation library (PyTorch Paszke et al. (2019)) for fair and exhaustive comparison. Regarding the anomaly detection model for Anomaly Prediction, we set the window size of 100, and used sliced predicted signals for obtaining the output of anomaly detection for experiments including all comparing methods.

B Analysis on Detecting Future Anomalies

To further investigate the effectiveness of our proposed A2P, we examined the anomaly scores of predicted signals, in which the time steps with higher anomaly scores are considered more likely to be anomalies. As shown in Figure 5, A2P exceeds in Anomaly Prediction with much higher anomaly scores in ground-truth anomaly time points depicted as red area, while our baseline, which is a naive combination of Patch-TST and Anomaly Transformer fails at Anomaly Prediction.

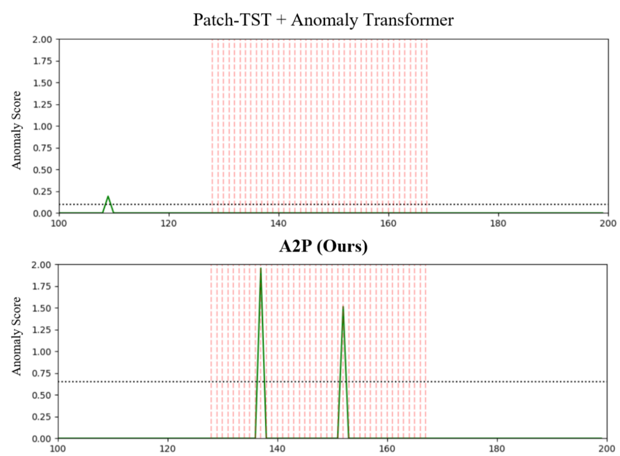

Figure 5: The anomaly score of the baseline, which is a naive combination of Patch-TST and Anomaly Transformer (top), and our proposed A2P (down) in MBA dataset.

C Hyperparameters.

Hyperparameter Sensitivity. To figure out the effect of various hyperparameters used in A2P, we examined the F1-Scores varying each hyperparameters as shown in Figure 6 . We conducted experiments on various λ from 0.1 to 0.9 with especially λ_{Intra} and λ_{Inter}, which are used to weigh each loss term in the objective function. Our proposed model A2P showed stable performance across various values of λ. Regarding the hyperparameters of Synthetic Anomaly Prompting, we examined the effect of various values of N for the number of anomaly pool, pool size, and L_z which is the length of a anomaly prompt. While our proposed A2P achieved the stable performance for N and L_z for three datasets, pool size of the anomaly prompt pool affects the performance. Specifically, the F1-Score on MBA degrades with bigger pool size, indicating that the selection of appropriate pool size considering the size of the dataset is needed to fully leverage the effectiveness of SAP. We also examined the influence of nh_{AFFN} which is the number of heads in Anomaly-Aware Forecasting Network. As shown in the last plot of Figure 6, our proposed A2P performs robustly.

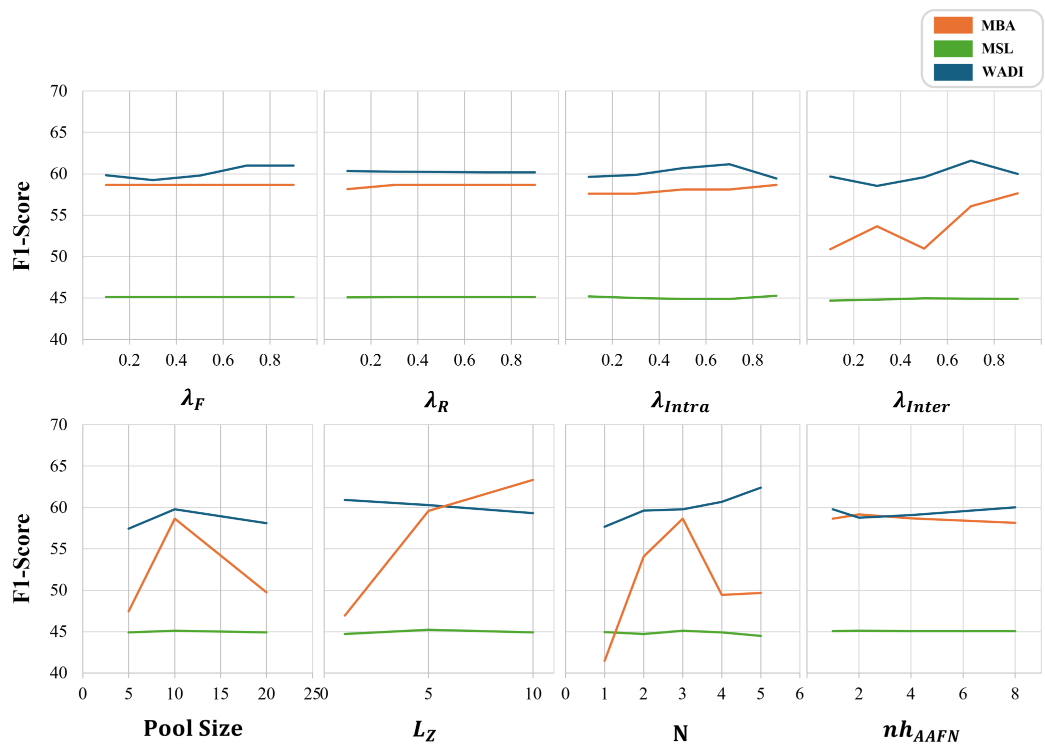

Figure 6: The results on various hyperparameter values.