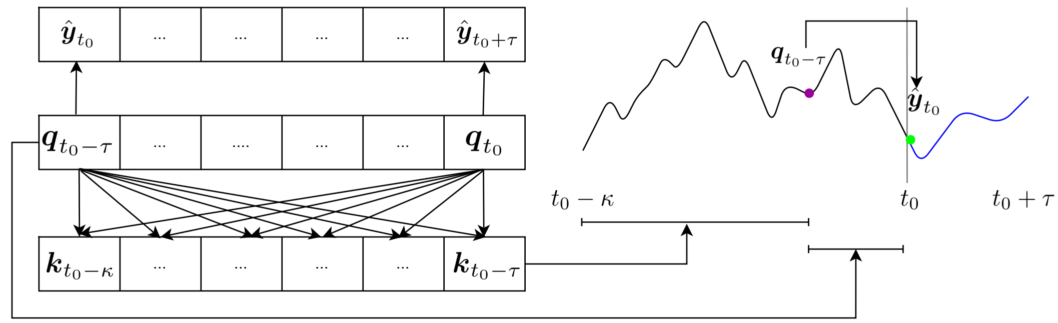

Figure 1: The queries $\{\boldsymbol{q}\}_{t=t_0-\tau}^{t_0}$ are the latent representations obtained from the decoder's self-attention, same-wise keys $\{\boldsymbol{k}\}_{t=t_0-\kappa}^{t_0-\tau}$ are the latent representations obtained from the encoder's self-attention. We make a multi-horizon prediction for each time step of interest from t_0 to $t_0+\tau$ by performing the cross-attention of the queries $\{\boldsymbol{q}\}_{t=t_0-\tau}^{t_0}$ and keys $\{\boldsymbol{k}\}_{t=t_0-\kappa}^{t_0-\tau}$. For example the prediction $\boldsymbol{y}_{t_0}$ is obtained by performing the cross-attention of query $\boldsymbol{q}_{t_0-\tau}$ and the keys $\{\boldsymbol{k}\}_{t=t_0-\kappa}^{t_0-\tau}$.

A Supplementary Explanation on How We Obtain Predictions

In Section 2.4, we describe our approach for generating forecasts for future τ time steps. This involves extracting the last τ values from the integration of two models: the forecasting model (referred to as Z_P) and the denoising model (referred to as Z_D). Our approach follows the encoder-decoder paradigm employed in transformers. However, unlike the conventional auto-regressive decoder method, where predictions are generated one element at a time, we employ a different strategy.

In an auto-regressive decoder, the initial prediction at time step t_0 is estimated based on the latent representation at time step t_0, denoted as the query at time step t_0, $\boldsymbol{q}_{t_0}$, along with the latent representations from time step $t_0-\tau$ to t_0 denoted as a sequence of keys $\{\boldsymbol{k}_t\}_{t=t_0-\kappa}^{t_0}$. In transformers with attention mechanism, the prediction $\hat{\boldsymbol{y}}_{t_0}$ is made by estimating the similarity between the query $\boldsymbol{q}_{t_0}$ and the keys $\{\boldsymbol{k}_t\}_{t=t_0-\kappa}^{t_0}$. Subsequent predictions are generated by recursively using the previous prediction as the query for the decoder. Each prediction is then added to the key sequence, expanding it for subsequent predictions.

This approach can be time consuming due to its sequential nature. For multi-horizon forecasting, we adopt a different approach. Following the self-attention mechanism in the encoder and decoder, the final predictions are generated using cross-attention. In this process, the queries are derived from the latent representations of the decoder, while the keys are derived from the latent representations of the encoder. However, in this approach, the decoder generates predictions at time step t_0 by utilizing the query at time step $t_0-\tau$ and the keys from time step $t_0-\kappa$ to $t_0-\tau$. Rather than iteratively providing predictions to the decoder, our objective is to make multi-horizon forecasts for each time step all at once. Consequently, to forecast at time step t_0+1, we employ the observation from time step $t_0-\tau+1$ as the query, and so forth. By adopting this method, we eliminate the iterative nature of the prediction process. Each forecast is made directly, without relying on previous predictions. In the main manuscript, we express this approach as $\{\hat{\boldsymbol{y}}\}_{t=t_0}^{t_0+\tau} = \{\boldsymbol{z}_{P_t}\}_{t=t_0-\tau}^{t_0} + \{\boldsymbol{z}_{D_t}\}_{t=t_0-\tau}^{t_0}$. Here, $\{\boldsymbol{z}_{P_t}\}_{t=t_0-\tau}^{t_0}$ and $\{\boldsymbol{z}_{D_t}\}_{t=t_0-\tau}^{t_0}$ refer to the predictions made by the decoder for future time steps from t_0 to $t_0+\tau$ with respect to the queries from time steps $t_0-\tau$ to t_0. On other words, given the query at time step $t_0-\tau$, we make a multi-step ahead prediction at time step τ, and given the query at time step $t_0-\tau+1$, we make the prediction at time step $\tau+1$. Please refer to Figure 1 for a depicted representation on how predictions are obtained.