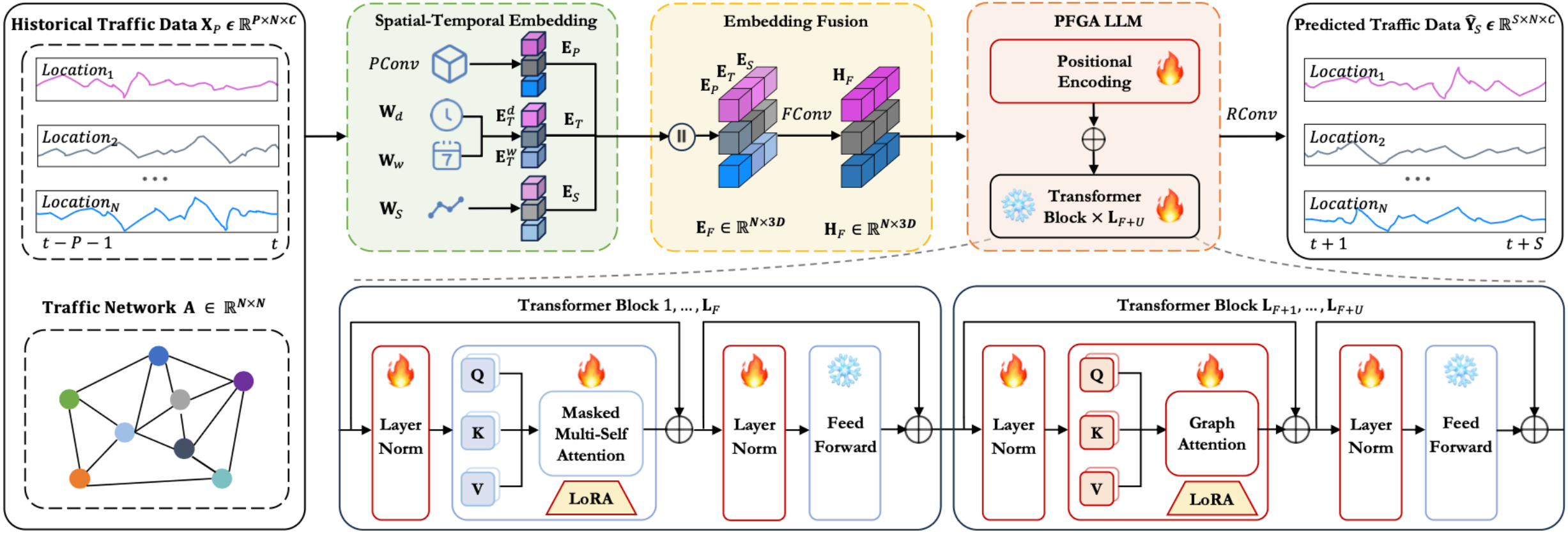

Fig. 1: ST-LLM+ Framework. ST-LLM+ modeling traffic data via a **Spatio-Temporal Embedding**. These embeddings are integrated uniformly by an **Embedding Fusion** layer. The **LoRA-augmented Partially Frozen Graph Attention (PFGA) LLMs** consist of $F + U$ layers: the first F layers employ the original multi-head attention mechanism, while the last U layers incorporate graph-based attention using an adjacency matrix as the attention mask. The multi-head attention and feed-forward layers in the first F layers are frozen, while multi-head attention layers are unfrozen and enhanced with LoRA in the last U layers. Finally, the output from the PFGA LLMs is passed through a regression layer to generate the traffic prediction results.