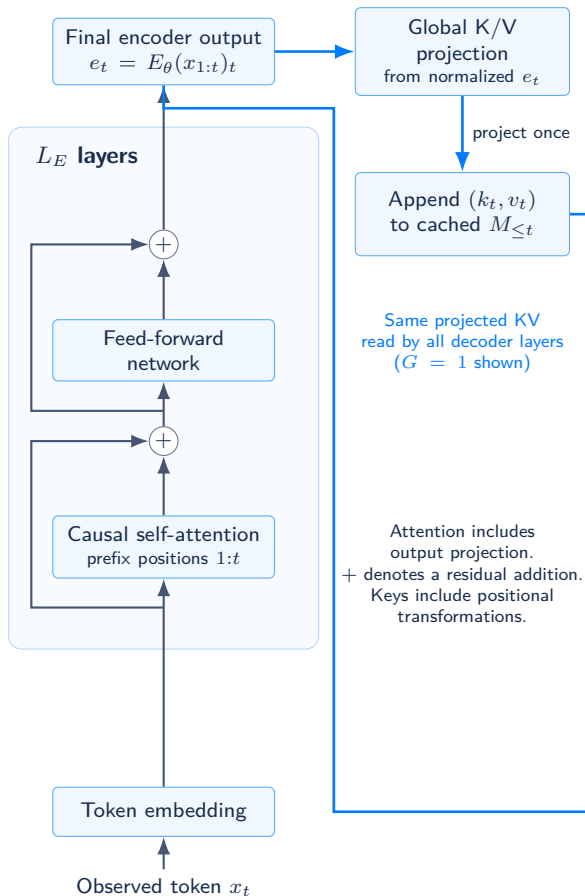


Causal encoder

Causal prefix $x_{1:t}$

Causal masking allows parallel encoder prefill across positions.



Recurrent decoder

One update at position t

