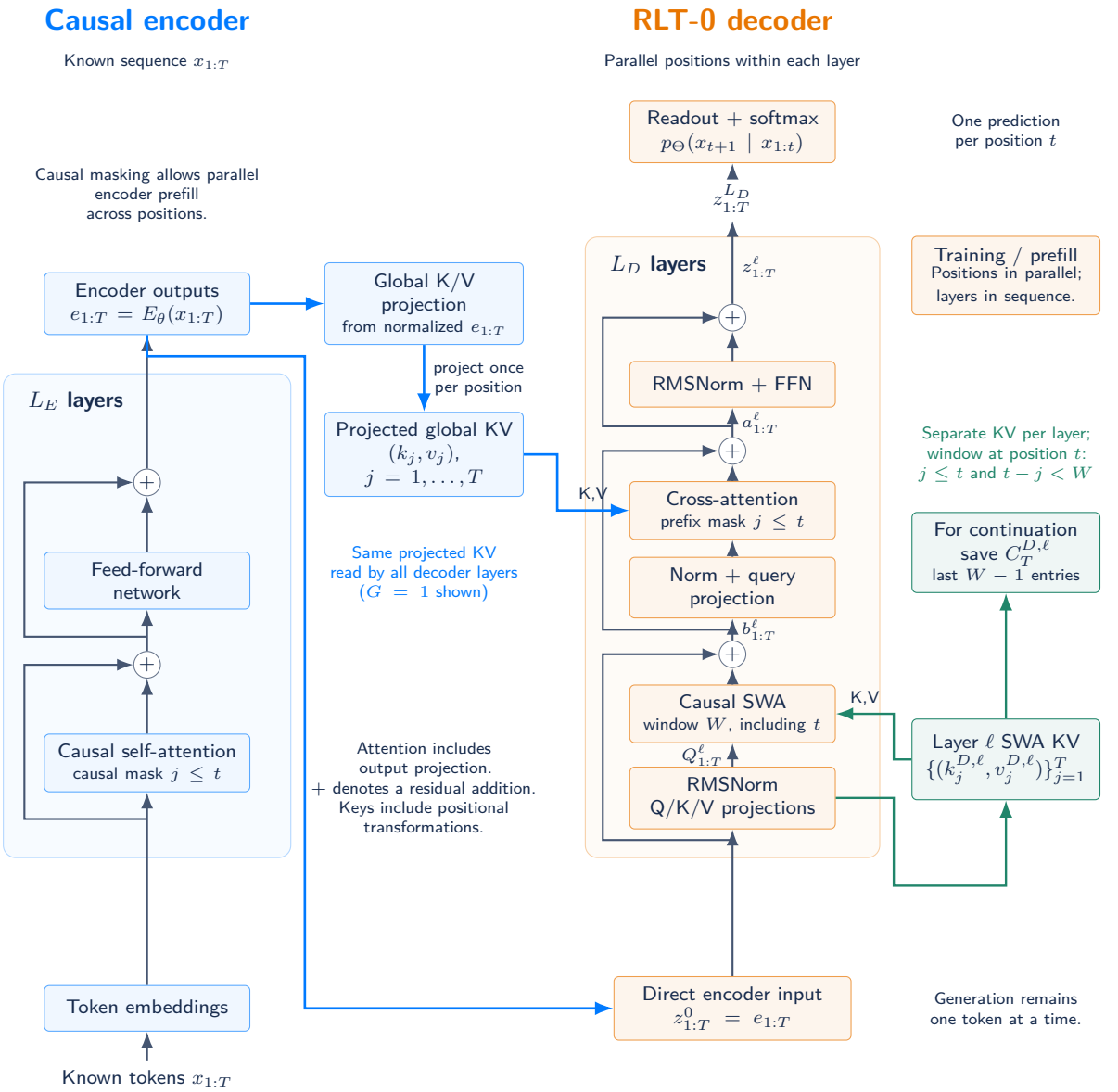
Causal encoder
Known sequence $x_{1:T}$
RLT-0 decoder
Parallel positions within each layer
Readout + softmax
$p_\Theta(x_{t+1} \mid x_{1:t})$
One prediction per position t
$z^{L_D}_{1:T}$
Causal masking allows parallel encoder prefill across positions.
L_D layers
$z^{\ell}_{1:T}$
Training / prefill
Positions in parallel; layers in sequence.
Encoder outputs
$e_{1:T} = E_\theta(x_{1:T})$
Global K/V projection
from normalized $e_{1:T}$
project once per position
RMSNorm + FFN
$a^{\ell}_{1:T}$
L_E layers
Projected global KV
(k_j, v_j), $j = 1, \ldots, T$
K,V
Separate KV per layer; window at position t: $j \le t$ and $t - j < W$
Cross-attention
prefix mask $j \le t$
Feed-forward network
Same projected KV read by all decoder layers ($G = 1$ shown)
Norm + query projection
For continuation save $C^{D,\ell}_T$
last $W - 1$ entries
$b^{\ell}_{1:T}$
Causal SWA
window W, including t
K,V
Causal self-attention
causal mask $j \le t$
Attention includes output projection.
+ denotes a residual addition.
Keys include positional transformations.
$Q^{\ell}_{1:T}$
Layer ℓ SWA KV
$\{(k^{D,\ell}_j, v^{D,\ell}_j)\}_{j=1}^{T}$
RMSNorm
Q/K/V projections
Token embeddings
Direct encoder input
$z^0_{1:T} = e_{1:T}$
Generation remains one token at a time.
Known tokens $x_{1:T}$