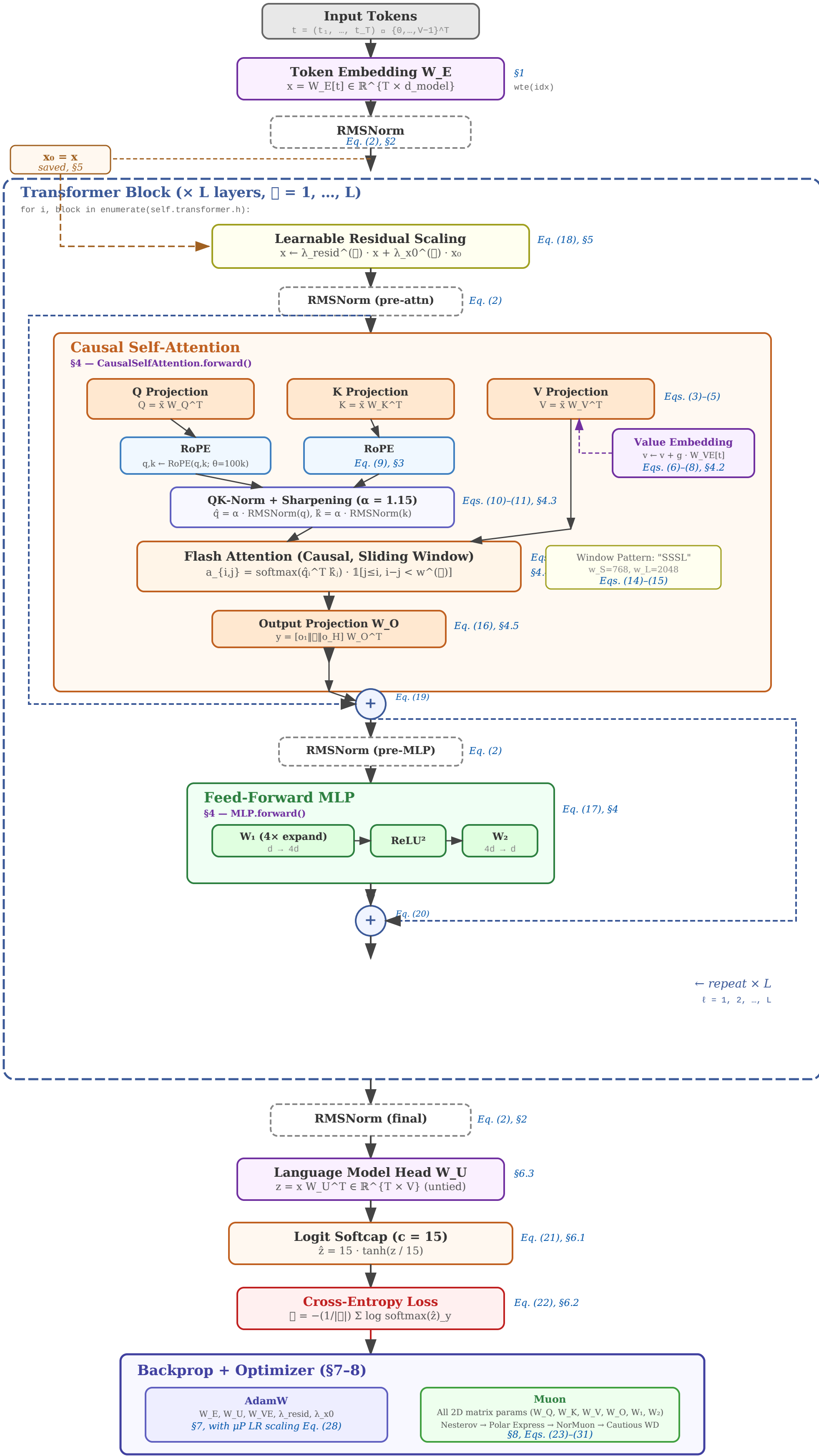


nanochat Architecture: Full Forward Pass

(Karpathy, 2025 — nanochat/gpt.py, forward() lines 398–434)



Legend

- Embedding layers
- Attention components
- MLP components
- RMSNorm (no params)
- Learnable scalars
- Loss computation
- Residual skip connection
- x_0 skip (input embedding reuse)

Key design choices vs. GPT-2:

- RoPE instead of learned positional embeddings
- RMSNorm (no params) instead of LayerNorm
- QK-Norm with $\alpha=1.15$ sharpening
- ReluSquared instead of GELU
- Untied $W_E \neq W_U$
- Logit softcap at $c=15$
- Sliding window "SSSL" pattern
- Value embeddings with gated residual (ResFormer)
- x_0 skip connection with learnable λ
- Muon optimizer for matrices, AdamW for rest