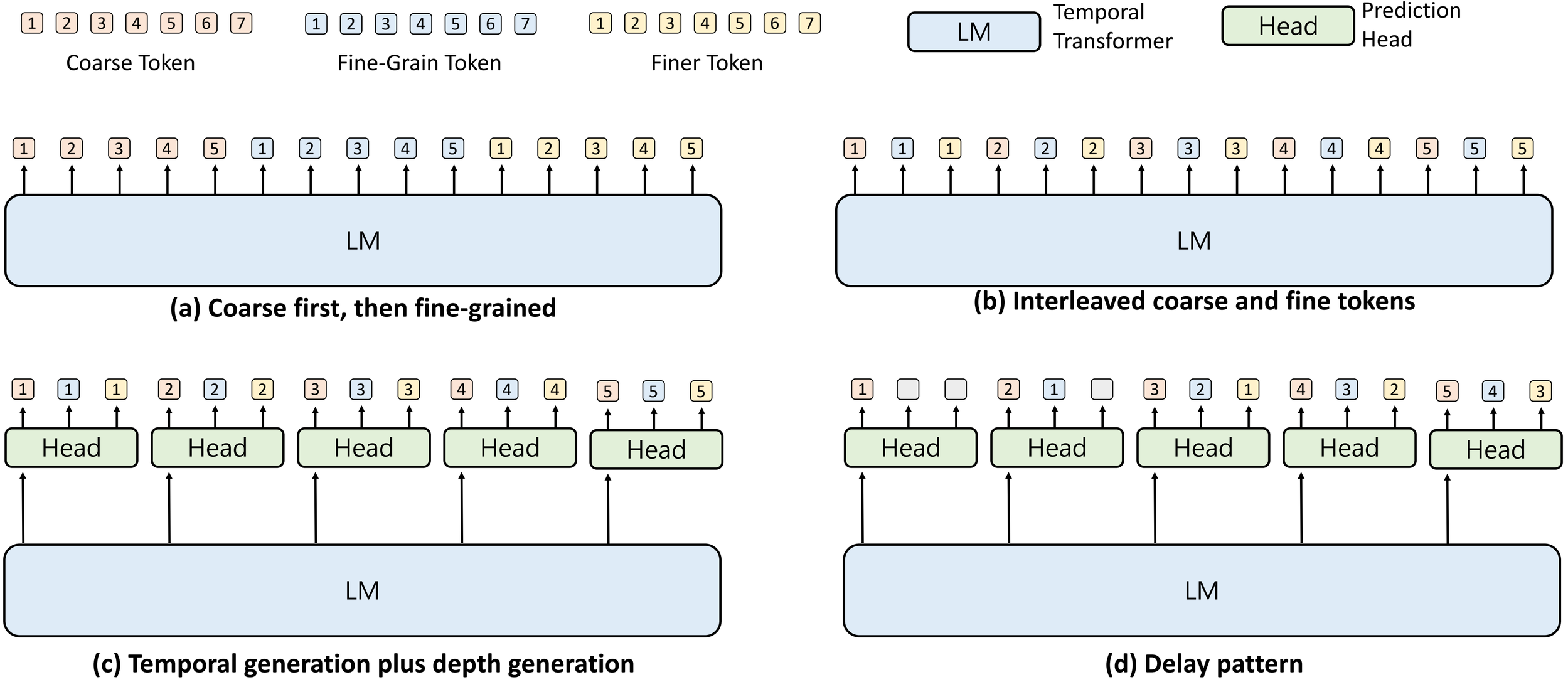
1 2 3 4 5 6 7
Coarse Token
1 2 3 4 5 6 7
Fine-Grain Token
1 2 3 4 5 6 7
Finer Token
LM
Temporal Transformer
Head
Prediction Head
1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
LM
(a) Coarse first, then fine-grained
1 1 1 2 2 2 3 3 3 4 4 4 5 5 5
LM
(b) Interleaved coarse and fine tokens
1 1 1 2 2 2 3 3 3 4 4 4 5 5 5
Head Head Head Head Head
LM
(c) Temporal generation plus depth generation
1 2 1 3 2 1 4 3 2 5 4 3
Head Head Head Head Head
LM
(d) Delay pattern