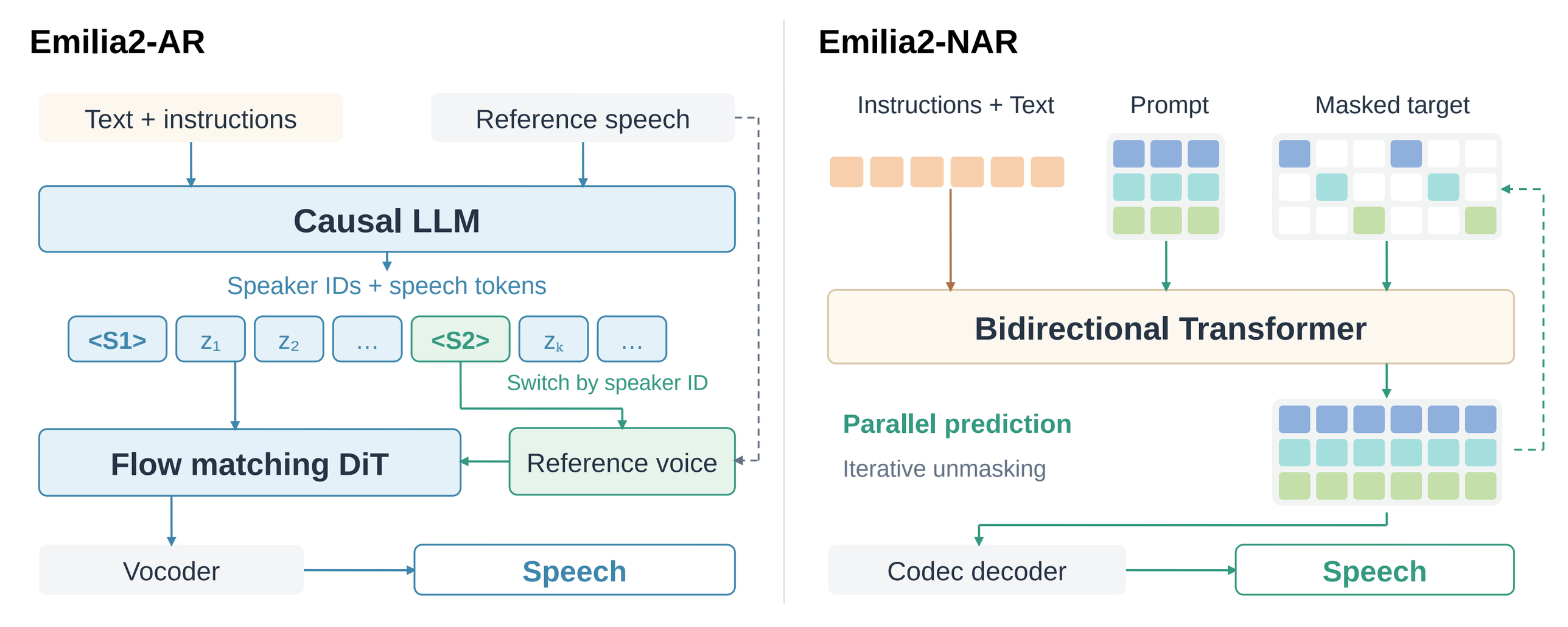

Emilia2-AR
Text + instructions
Reference speech
Causal LLM
Speaker IDs + speech tokens
<S1>
z1
z2
…
<S2>
zk
…
Switch by speaker ID
Flow matching DiT
Reference voice
Vocoder
Speech
Emilia2-NAR
Instructions + Text
Prompt
Masked target
Bidirectional Transformer
Parallel prediction
Iterative unmasking
Codec decoder
Speech