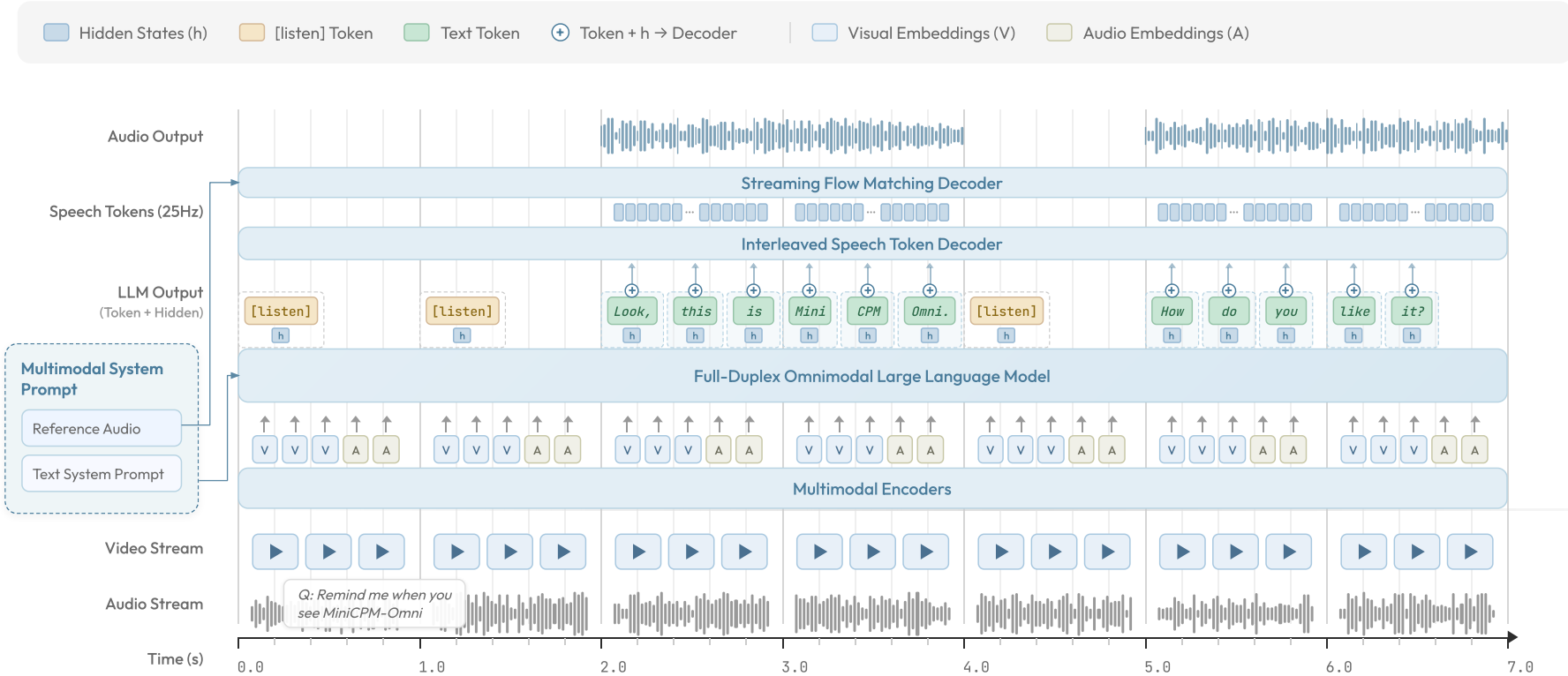
Hidden States (h)
[listen] Token
Text Token
Token + h → Decoder
Visual Embeddings (V)
Audio Embeddings (A)
Audio Output
Speech Tokens (25Hz)
LLM Output
(Token + Hidden)
Multimodal System Prompt
Reference Audio
Text System Prompt
Video Stream
Audio Stream
Time (s)
Streaming Flow Matching Decoder
Interleaved Speech Token Decoder
Full-Duplex Omnimodal Large Language Model
Multimodal Encoders
[listen]
[listen]
Look,
this
is
Mini
CPM
Omni.
[listen]
How
do
you
like
it?
Q: Remind me when you see MiniCPM-Omni
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0