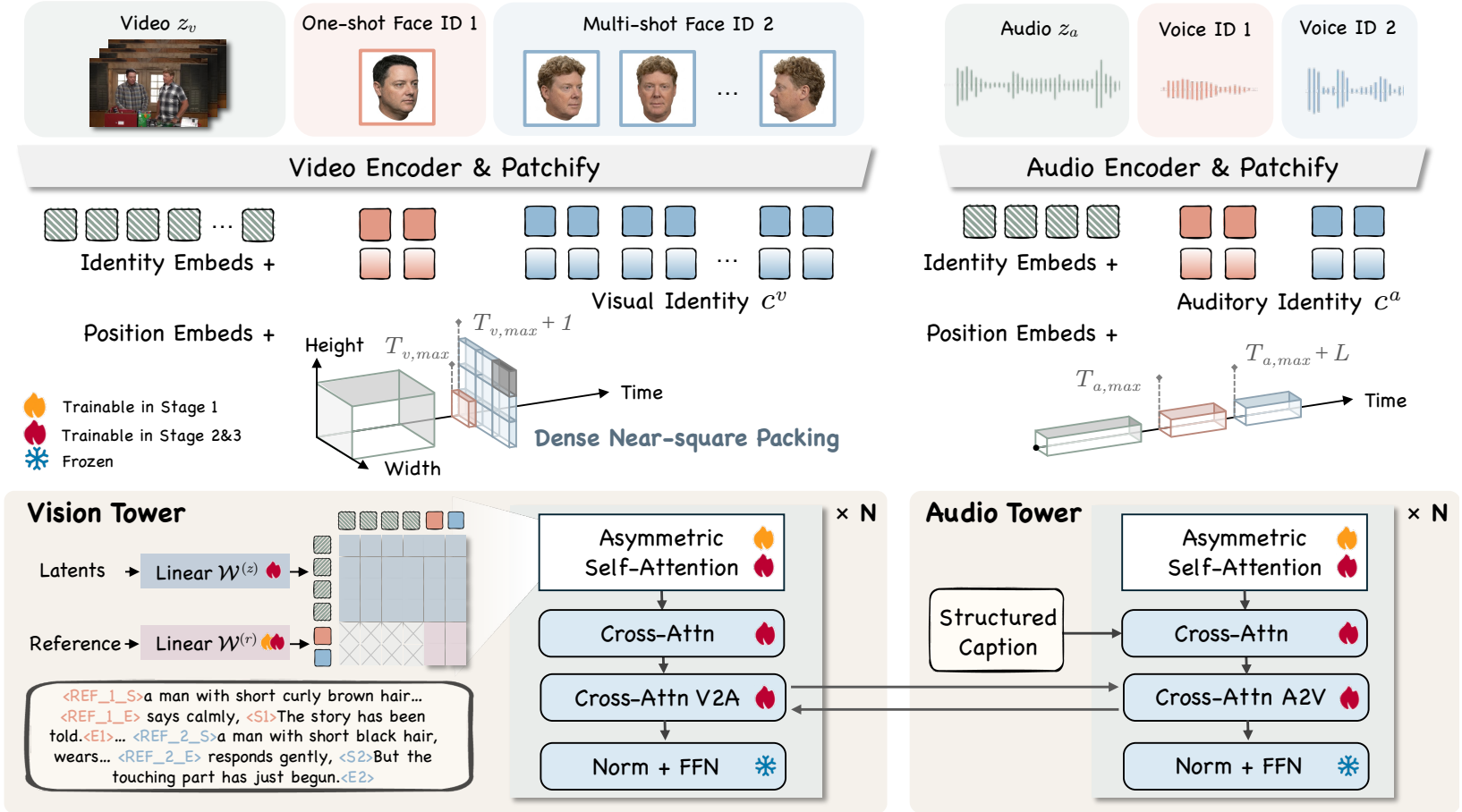
Video z_v
One-shot Face ID 1
Multi-shot Face ID 2
Audio z_a
Voice ID 1
Voice ID 2
Video Encoder & Patchify
Audio Encoder & Patchify
Identity Embeds +
Visual Identity c^v
Position Embeds +
Height
$T_{v,max}$
$T_{v,max}+1$
Time
Width
Dense Near-square Packing
Trainable in Stage 1
Trainable in Stage 2&3
Frozen
Identity Embeds +
Auditory Identity c^a
Position Embeds +
$T_{a,max}$
$T_{a,max}+L$
Time
Vision Tower
Latents
Linear $\mathcal{W}^{(z)}$
Reference
Linear $\mathcal{W}^{(r)}$
<REF_1_S>a man with short curly brown hair... <REF_1_E> says calmly, <S1>The story has been told.<E1>... <REF_2_S>a man with short black hair, wears... <REF_2_E> responds gently, <S2>But the touching part has just begun.<E2>
Asymmetric Self-Attention
Cross-Attn
Cross-Attn V2A
Norm + FFN
× N
Audio Tower
Structured Caption
Asymmetric Self-Attention
Cross-Attn
Cross-Attn A2V
Norm + FFN
× N