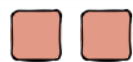


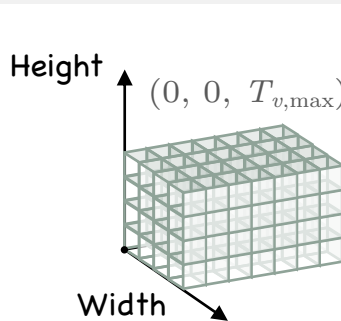
Video Encoder & Patchify



Identity Embeds



Position Embeds



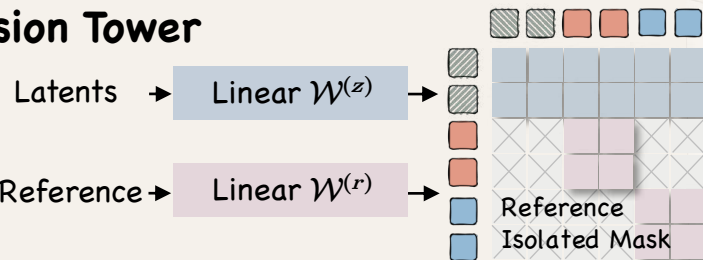
Spatial Packing

Time

Temporal Expansion

- Trainable in Stage 1
- Trainable in Stage 2&3
- Frozen

Vision Tower



<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>

Structured Caption

Asymmetric
Self-Attention

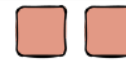
$\times N$

Cross-Attn

Cross-Attn A2V

Norm + FFN

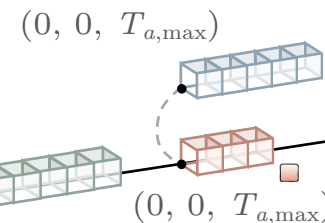
Audio Encoder & Patchify



Identity Embeds



Position Embeds



Time

Auditory Identity C^a

Audio Tower

Structured
Caption

Asymmetric
Self-Attention

$\times N$

Cross-Attn

Cross-Attn V2A

Norm + FFN