

noise Semantic tokens  $x_t$ Xunmasked D Projection Transformer 3D emotion representation solving ODE  $s_{\theta}(z,t)$ audio (b) Stage 2 Prosody-guided zero-shot synthesis

Mask

prompt

prompt

(a) Stage 1 Flexible control of expressiveness via keyframes