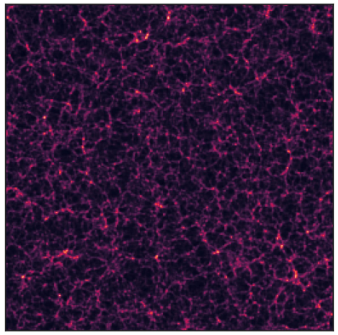
(a) High-level architecture



1 × 256 × 256
(channels × height × width)

Stage 1

64 × 64 × 64

Stage 2

192 × 32 × 32

Stage 3

384 × 16 × 16

MLP head

$\mu_{1,j}$, $\sigma_{1,j}$

$\mu_{2,j}$, $\sigma_{2,j}$

(b) Stage "x"

Convolutional Token Embedding

2D convolution operation

tokens

Convolutional Transformer Block

× N