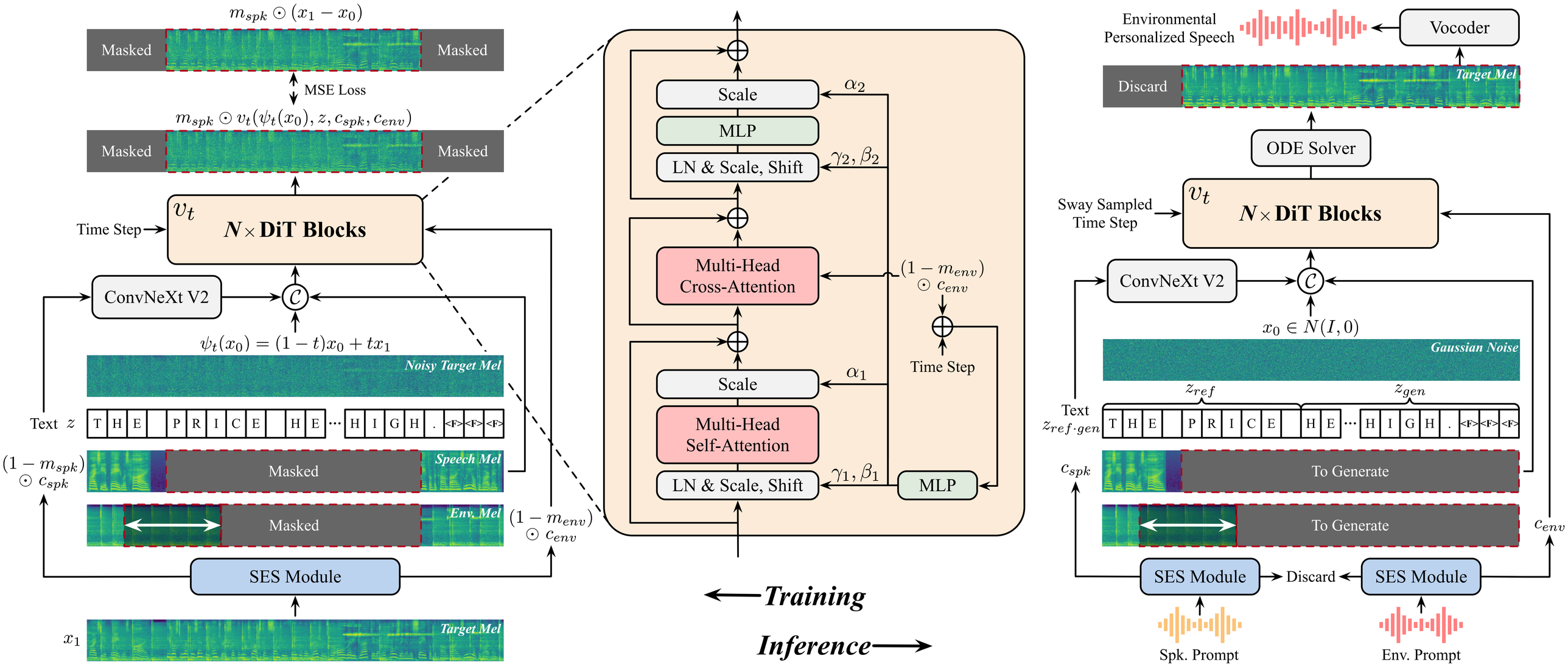
$m_{spk} \odot (x_1 - x_0)$
Masked
Masked
MSE Loss
$m_{spk} \odot v_t(\psi_t(x_0), z, c_{spk}, c_{env})$
Masked
Masked
Time Step
v_t
$N \times$ DiT Blocks
ConvNeXt V2
$\mathcal{C}$
$\psi_t(x_0) = (1-t)x_0 + tx_1$
Noisy Target Mel
Text z
T H E P R I C E H E ··· H I G H . <F> <F> <F>
$(1 - m_{spk}) \odot c_{spk}$
Masked
Speech Mel
Masked
Env. Mel
$(1 - m_{env}) \odot c_{env}$
SES Module
x_1
Target Mel
Scale
MLP
LN & Scale, Shift
α_2
γ_2, β_2
Multi-Head Cross-Attention
$(1 - m_{env}) \odot c_{env}$
Time Step
Scale
α_1
Multi-Head Self-Attention
LN & Scale, Shift
γ_1, β_1
MLP
Training
Inference
Environmental Personalized Speech
Vocoder
Discard
Target Mel
ODE Solver
Sway Sampled Time Step
v_t
$N \times$ DiT Blocks
ConvNeXt V2
$\mathcal{C}$
$x_0 \in N(I, 0)$
Gaussian Noise
z_{ref}
z_{gen}
Text $z_{ref \cdot gen}$
T H E P R I C E H E ··· H I G H . <F> <F> <F>
c_{spk}
To Generate
To Generate
c_{env}
SES Module
Discard
SES Module
Spk. Prompt
Env. Prompt