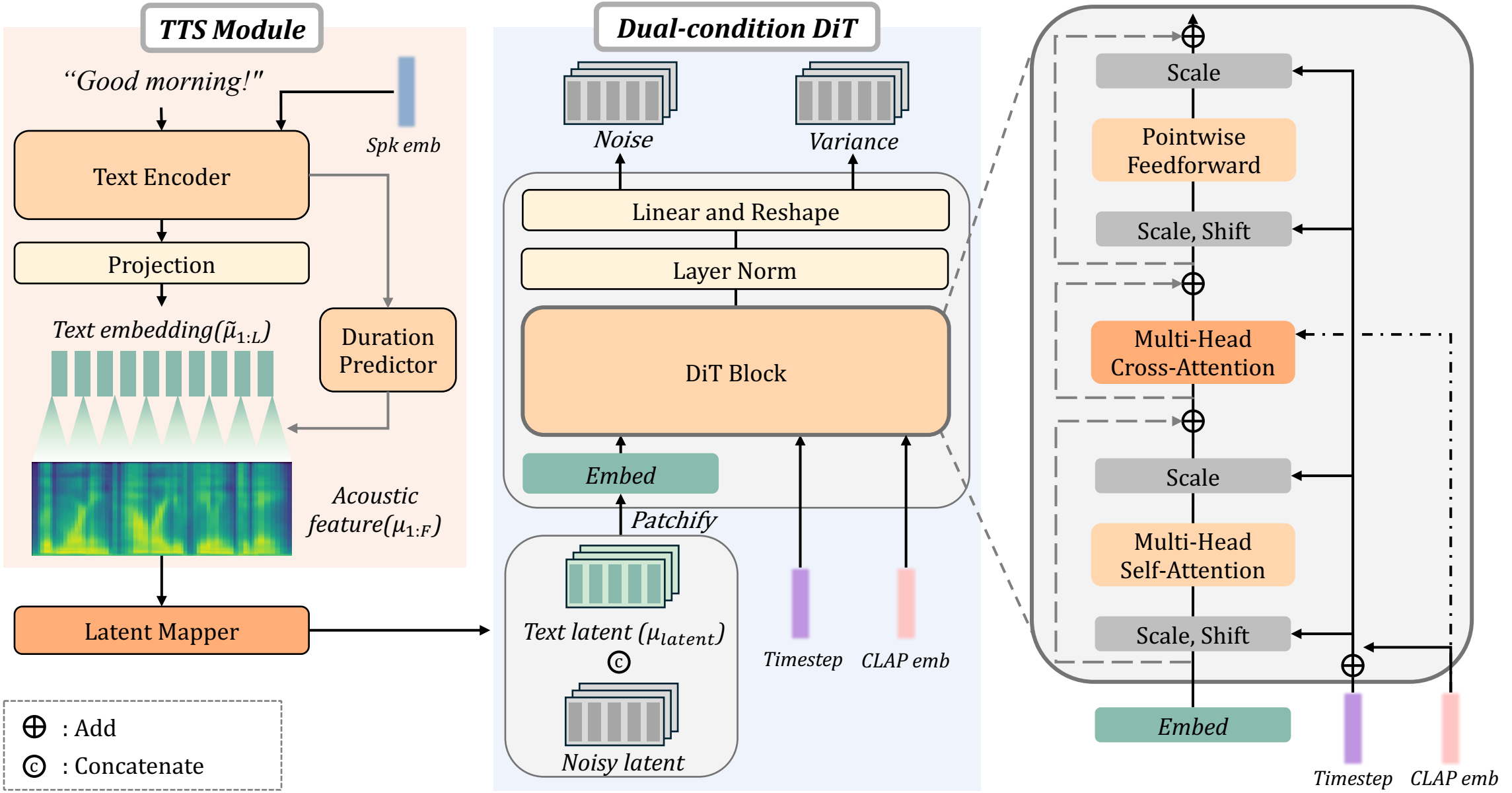
TTS Module
"Good morning!"
Spk emb
Text Encoder
Projection
Text embedding($\tilde{\mu}_{1:L}$)
Duration Predictor
Acoustic feature($\mu_{1:F}$)
Latent Mapper
Dual-condition DiT
Noise
Variance
Linear and Reshape
Layer Norm
DiT Block
Embed
Patchify
Text latent (μ_{latent})
Noisy latent
Timestep
CLAP emb
Scale
Pointwise Feedforward
Scale, Shift
Multi-Head Cross-Attention
Scale
Multi-Head Self-Attention
Scale, Shift
Embed
Timestep
CLAP emb
: Add
: Concatenate