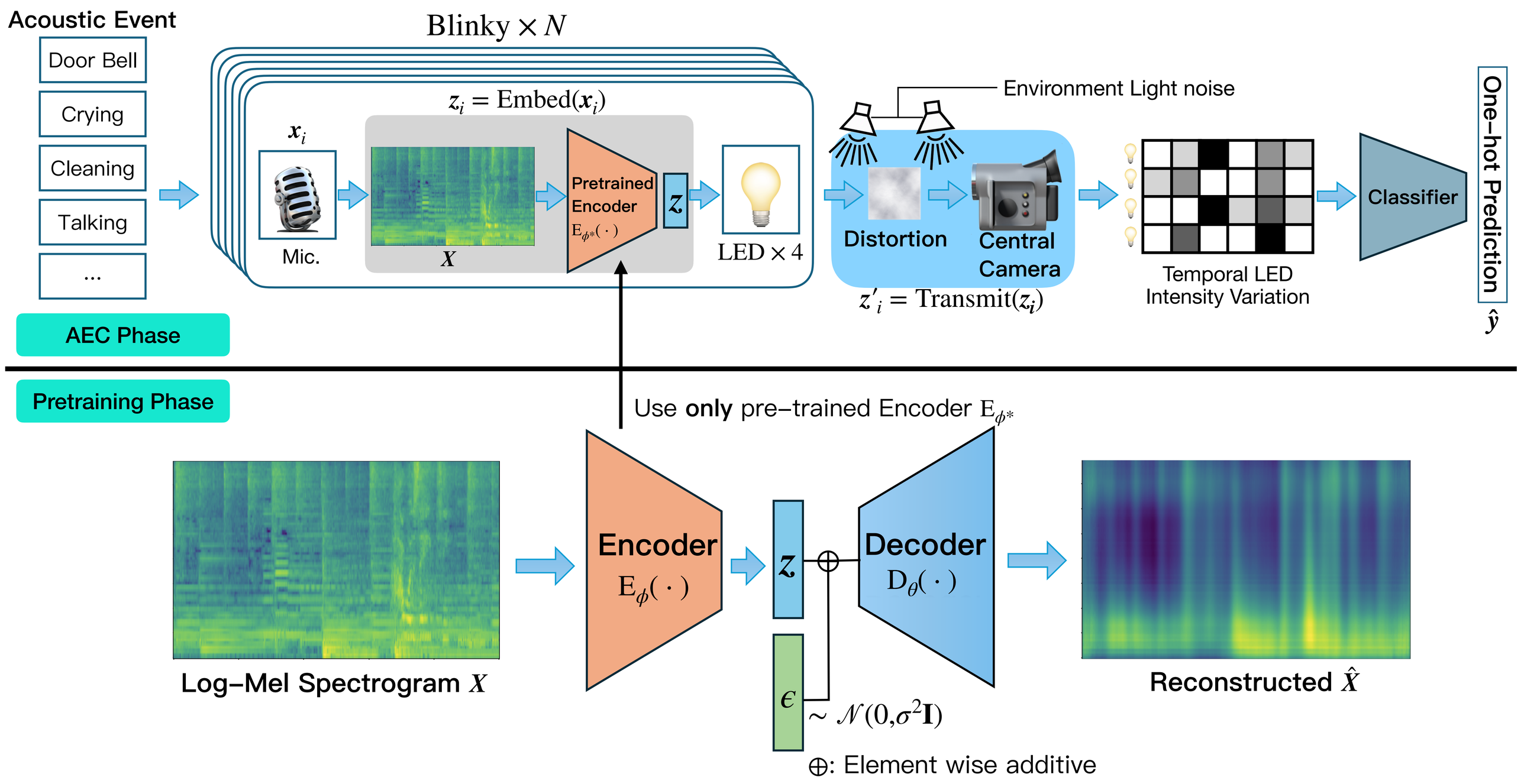
Acoustic Event
Door Bell
Crying
Cleaning
Talking
...
Blinky × N
$z_i = \mathrm{Embed}(x_i)$
x_i
Mic.
X
Pretrained Encoder
$\mathrm{E}_{\phi^*}(\cdot)$
z
LED × 4
Environment Light noise
Distortion
Central Camera
$z'_i = \mathrm{Transmit}(z_i)$
Temporal LED Intensity Variation
Classifier
One–hot Prediction
$\hat{y}$
AEC Phase
Pretraining Phase
Use only pre–trained Encoder E_{ϕ^*}
Log–Mel Spectrogram X
Encoder
$\mathrm{E}_{\phi}(\cdot)$
z
ϵ
$\sim \mathcal{N}(0,\sigma^2\mathbf{I})$
Decoder
$\mathrm{D}_{\theta}(\cdot)$
Reconstructed $\hat{X}$
⊕: Element wise additive