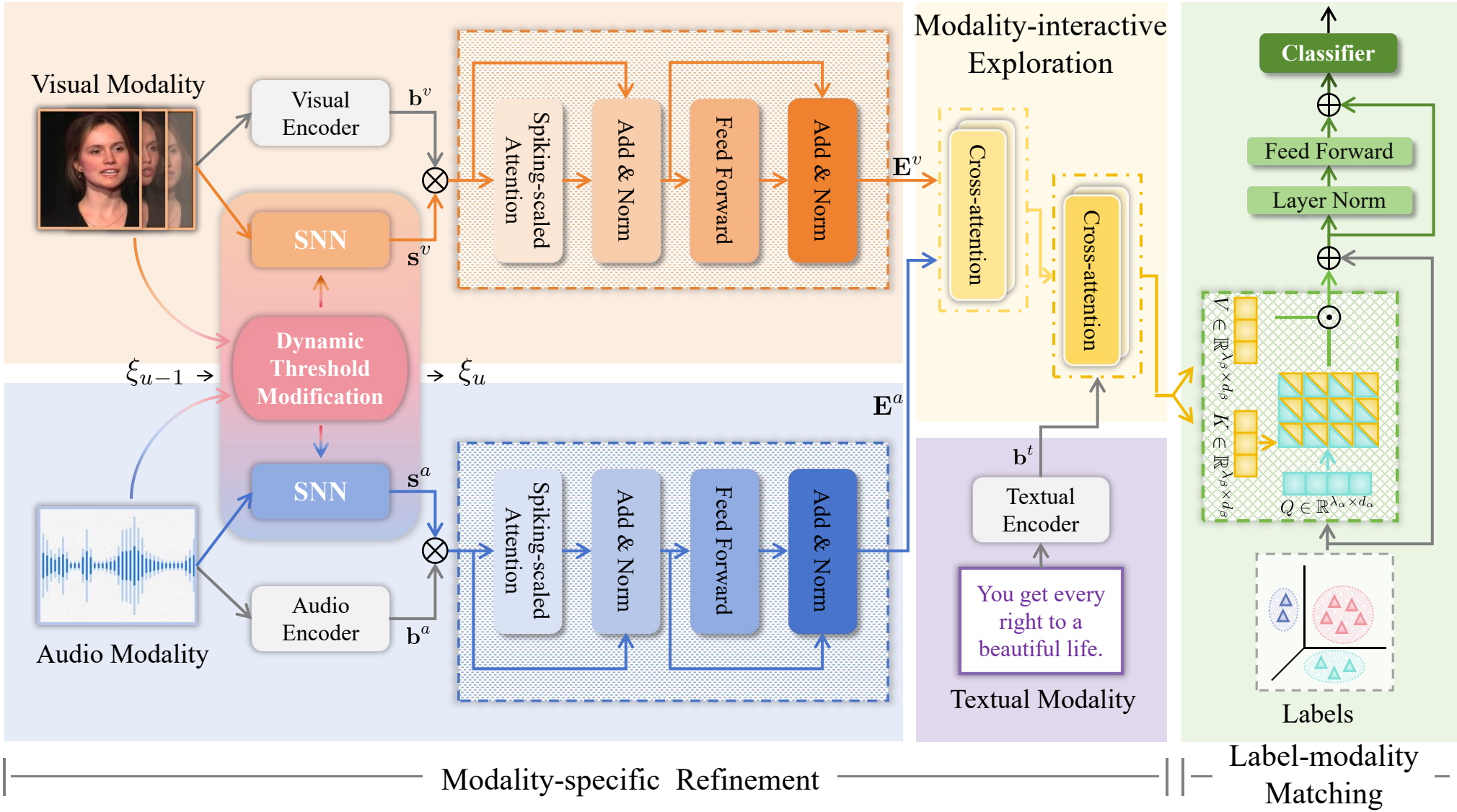
Visual Modality
Visual Encoder
$\mathbf{b}^v$
SNN
$\mathbf{s}^v$
Spiking-scaled Attention
Add & Norm
Feed Forward
Add & Norm
$\mathbf{E}^v$
Modality-interactive Exploration
Cross-attention
Cross-attention
Classifier
Feed Forward
Layer Norm
$V \in \mathbb{R}^{\lambda_\beta \times d_\beta}$
$K \in \mathbb{R}^{\lambda_\beta \times d_\beta}$
$Q \in \mathbb{R}^{\lambda_\alpha \times d_\alpha}$
Dynamic Threshold Modification
ξ_{u-1}
ξ_u
$\mathbf{E}^a$
SNN
$\mathbf{s}^a$
Audio Modality
Audio Encoder
$\mathbf{b}^a$
Spiking-scaled Attention
Add & Norm
Feed Forward
Add & Norm
$\mathbf{b}^t$
Textual Encoder
You get every right to a beautiful life.
Textual Modality
Labels
Modality-specific Refinement
Label-modality Matching