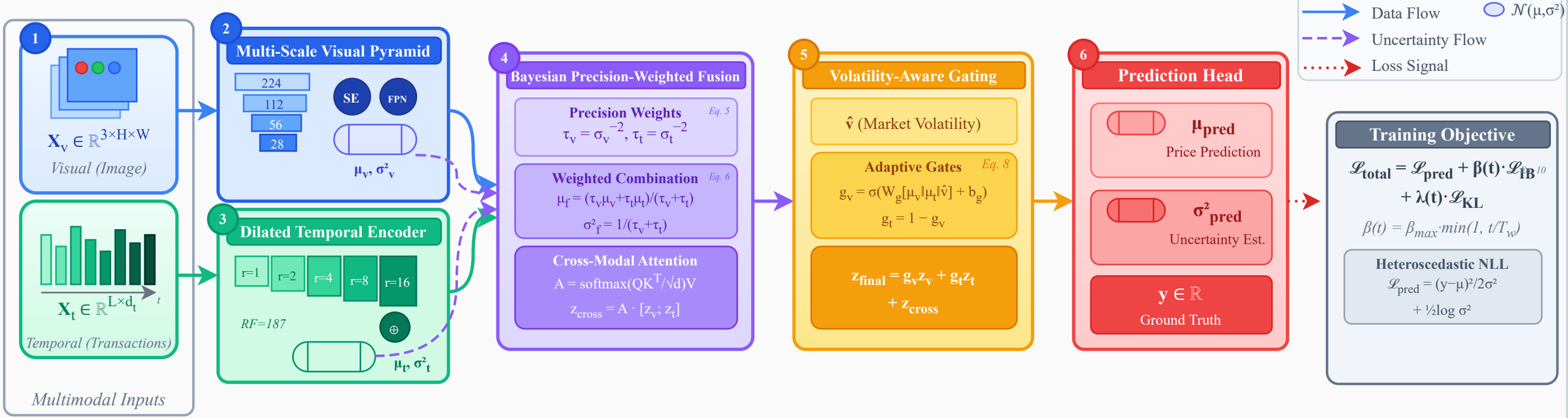
(a) Complete UIBFuse Pipeline



(b) Multi-Scale Visual Pyramid

Level 1: 224×224, C=64

Level 2: 112×112, C=128

Level 3: 56×56, C=256

Level 4: 28×28, C=512

SE-Net Attention
Squeeze: GAP → FC(C/r)
Excitation: FC(C) → σ
Scale: x · σ(s)

FPN Multi-Scale Fusion
Top-down pathway
Lateral connections
Element-wise addition

Output: $\mu_v \in \mathbb{R}^{256}$, $\sigma^2_v \in \mathbb{R}^{256}_{>0}$

Spectral: $S(f) \propto f^{-2.1}$ → *4 levels capture 95% info*

(c) Dilated Temporal Encoder

$X_t \in \mathbb{R}^{L\times 64}$

$RF = 1 + \Sigma_i(k-1)\cdot r_i =$ **187 steps**

Dilated Convolution Stack (Octave-based Rates)

r=1	r=2	r=4	r=8	r=16
(2^0)	(2^1)	(2^2)	(2^3)	(2^4)

Transformer Layers (H=16 heads, d_{model}=256)

Output: $\mu_t \in \mathbb{R}^{256}$, $\sigma^2_t \in \mathbb{R}^{256}_{>0}$

(d) Information-Theoretic Derivations

Information Bottleneck Objective

$\mathcal{L}_{IB} = -I(Z; Y) + \beta \cdot I(Z; X_v, X_t)$

① **Latent Dimension (d_z)**

$d_z \geq 2\cdot I((V,T);Y) / (\beta\cdot\log(2\pi e))$

$I \approx 4.3$ nats, $\beta = 0.1$ → $\mathbf{d_z = 256}$

② **Pyramid Scales**

$S(f) \propto f^{-\alpha}$, $\alpha \approx 2.1$ → 95% info at **{224, 112, 56, 28}**

③ **Attention Heads (H)**

Spectral clustering → 10-20 modes → **H = 16**