Gated Attention CEM（与原报告一致）

输入批次
特征 Z（割点输出）+ 标签 Y

按标签分组
得到每个类的样本矩阵 X_c

LayerNorm
$\hat{X}_c = \text{LayerNorm}(X_c)$

GatedAttentionPooling（报告 §2）
$V = \tanh(W_V \hat{X}_c),\ U = \sigma(W_U \hat{X}_c)$
$\text{logits} = W_w(V \odot U),\ a = \text{softmax}(\text{logits})$

加权均值
$\mu_c = \Sigma_m\, a_m\, \hat{x}_m$

加权方差
$\sigma_c^2 = \Sigma_m\, a_m\, (\hat{x}_m - \mu_c)^2$
clamp 至 eps，避免 log0

阈值化 log-variance
$\log\sigma^2 = \log(\sigma^2 + \gamma)$
$\text{ce\_sur} = \max(0, \log\sigma^2 - \log\sigma^2_{thr})$

类级输出
$L_c = \text{mean}(\text{ce\_sur}),\ MSE_c = \text{mean}\|\hat{x} - \mu\|^2$
权重 $w_c = M/B$

聚合
$\text{rob\_loss} = \Sigma_c\, w_c\, L_c$
$\text{intra\_mse} = \Sigma_c\, w_c\, MSE_c$

训练整合
· rob_loss × attention_loss_scale
· 先反向 rob_loss，缓存梯度
· 再反向主任务并合并梯度
· NaN/Inf → rob_loss 置零