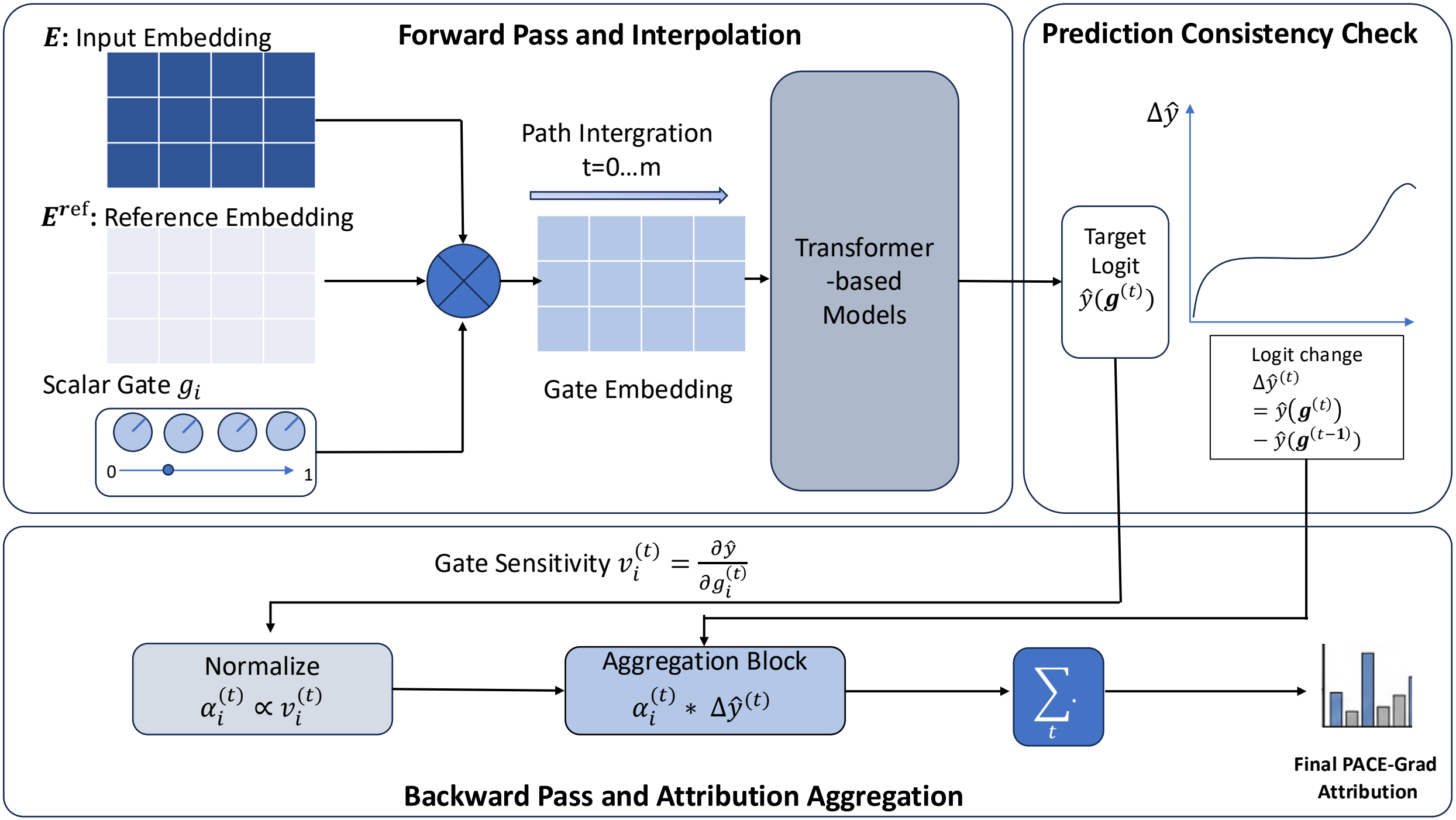
E: Input Embedding
E^{ref}: Reference Embedding
Scalar Gate g_i
0
1
Forward Pass and Interpolation
Path Intergration
t=0…m
Gate Embedding
Transformer
-based
Models
Prediction Consistency Check
$\Delta\hat{y}$
Target
Logit
$\hat{y}(\boldsymbol{g}^{(t)})$
Logit change
$\Delta\hat{y}^{(t)}$
$= \hat{y}(\boldsymbol{g}^{(t)})$
$- \hat{y}(\boldsymbol{g}^{(t-1)})$
Gate Sensitivity $v_i^{(t)} = \frac{\partial \hat{y}}{\partial g_i^{(t)}}$
Normalize
$\alpha_i^{(t)} \propto v_i^{(t)}$
Aggregation Block
$\alpha_i^{(t)} * \Delta\hat{y}^{(t)}$
$\sum_t$
Final PACE-Grad
Attribution
Backward Pass and Attribution Aggregation