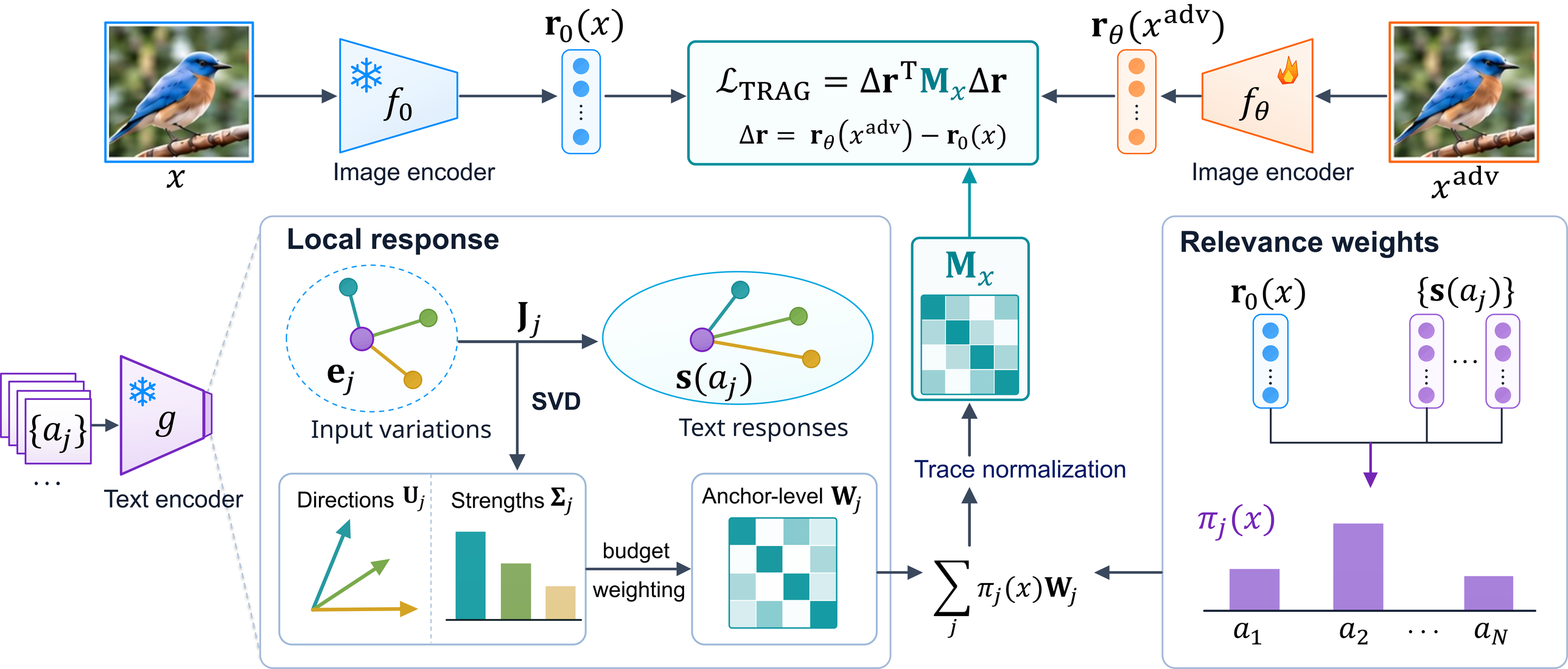

x
f_0
Image encoder
$\mathbf{r}_0(x)$
$\mathcal{L}_{\text{TRAG}} = \Delta\mathbf{r}^{\text{T}}\mathbf{M}_x\Delta\mathbf{r}$
$\Delta\mathbf{r} = \mathbf{r}_\theta(x^{\text{adv}}) - \mathbf{r}_0(x)$
$\mathbf{r}_\theta(x^{\text{adv}})$
f_θ
Image encoder
x^{adv}
$\{a_j\}$
g
Text encoder
Local response
$\mathbf{e}_j$
Input variations
$\mathbf{J}_j$
SVD
$\mathbf{s}(a_j)$
Text responses
Directions $\mathbf{U}_j$
Strengths $\mathbf{\Sigma}_j$
budget weighting
Anchor-level $\mathbf{W}_j$
$\sum_j \pi_j(x)\mathbf{W}_j$
Trace normalization
$\mathbf{M}_x$
Relevance weights
$\mathbf{r}_0(x)$
$\{\mathbf{s}(a_j)\}$
$\pi_j(x)$
a_1
a_2
$\cdots$
a_N