Zero-shot Guidance

a photo of {cls}, {$d_0$}

a photo of {cls}, {$d_3$}

Text Encoder

$w_{0,1}^{sem}$ $w_{0,2}^{sem}$ ... $w_{0,c}^{sem}$

$w_{3,1}^{sem}$ $w_{3,2}^{sem}$ ... $w_{3,c}^{sem}$

$w_1^{zst}$ $w_2^{zst}$ ... $w_c^{zst}$

0 1 ... 0

One − hot label

$p_1^{zs}$ $p_2^{zs}$ ... $p_{Nc}^{zs}$

Image Encoder

$w_1^{zsv}$ $w_2^{zsv}$ ... $w_c^{zsv}$

$\alpha$

$L_{zvg}$ $L_{sg}^{str}$ $L_{sg}^{weak}$ $L_{zpg}$ $L_{ce}$

$v_1^{vis}$ $v_2^{vis}$ ... $v_m^{vis}$

Image Encoder

$w_1^{vis}$ $w_2^{vis}$ ... $w_c^{vis}$

$\text{prompt}_{\text{layer}_i}^{\text{vis}}$

$v_1^{str}$ $v_2^{str}$ ... $v_m^{str}$ {cls}

$v_1^{weak}$ $v_2^{weak}$ ... $v_m^{weak}$ {cls}

Text Encoder

$w_1^{str}$ $w_2^{str}$ ... $w_c^{str}$

$w_1^{weak}$ $w_2^{weak}$ ... $w_c^{weak}$

$\text{prompt}_{\text{layer}_i}^{\text{text}}$

$l_1^{str}$ $l_2^{str}$ ... $l_{Nc}^{str}$

$l_1^{weak}$ $l_2^{weak}$ ... $l_{Nc}^{weak}$

$p_1^{fin}$ $p_2^{fin}$ ... $p_{Nc}^{fin}$

Dual branch&modal Learning

$\xi_1$ $\xi_2$ $\xi_3$ ... $\xi$

training epoch model parameter

$$\sum_{t=1}^{T} \varepsilon_t \cdot \xi_t, \quad \varepsilon_t \sim \mathcal{N}(\mu, \sigma^2)$$

Epoch Weight Averaging

$\xi_{\text{final}}$

EWA Fusion