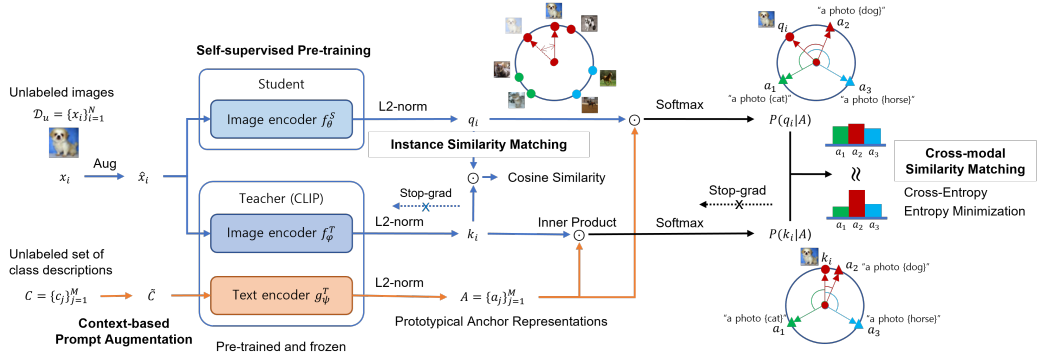

Self-supervised Pre-training
Unlabeled images
$\mathcal{D}_u = \{x_i\}_{i=1}^N$
Aug
x_i
$\hat{x}_i$
Student
Image encoder f_θ^S
L2-norm
q_i
Instance Similarity Matching
Cosine Similarity
Stop-grad
Teacher (CLIP)
Image encoder f_ψ^T
L2-norm
k_i
Inner Product
Unlabeled set of class descriptions
$C = \{c_j\}_{j=1}^M$
$\tilde{C}$
Text encoder g_ψ^T
L2-norm
$A = \{a_j\}_{j=1}^M$
Prototypical Anchor Representations
Context-based Prompt Augmentation
Pre-trained and frozen
Softmax
$P(q_i|A)$
Stop-grad
Softmax
$P(k_i|A)$
"a photo (dog)"
"a photo (cat)"
"a photo (horse)"
Cross-modal Similarity Matching
Cross-Entropy
Entropy Minimization