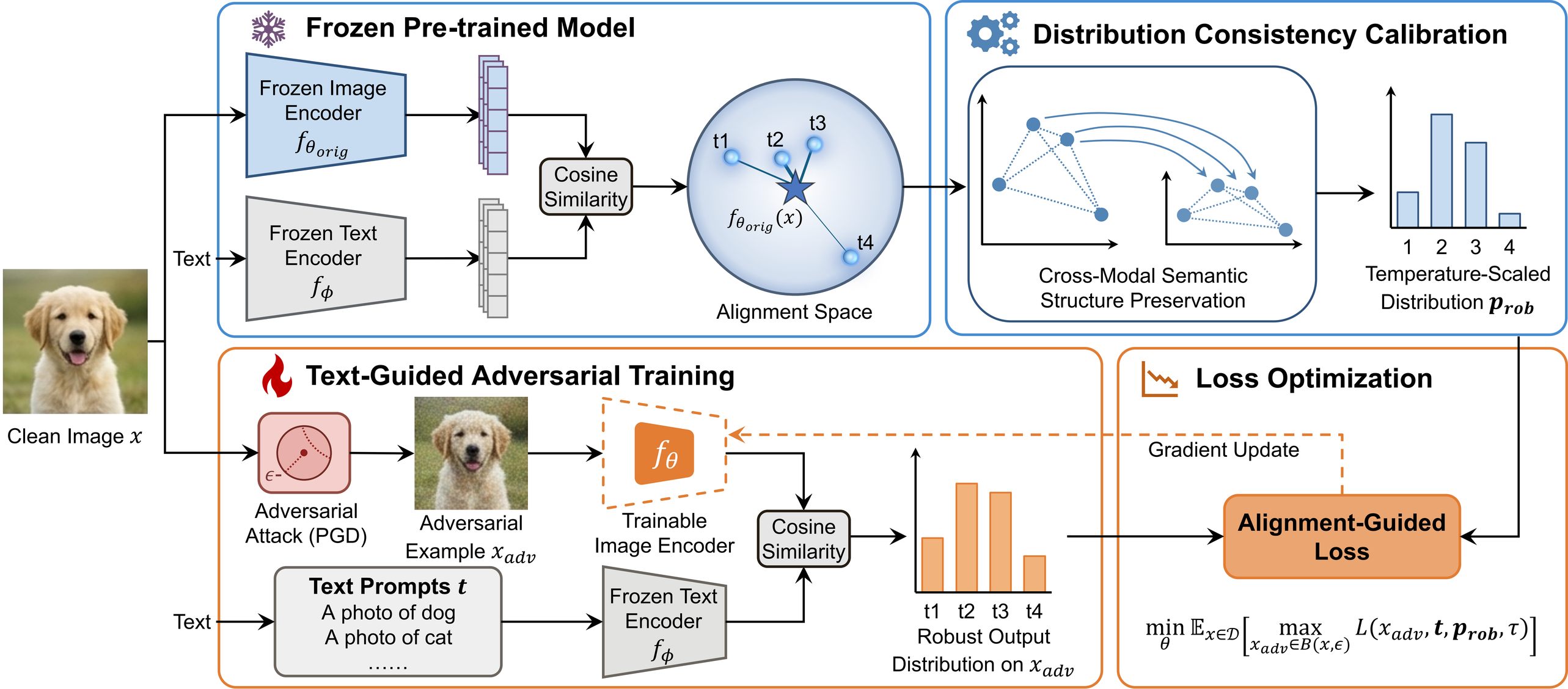

Frozen Pre-trained Model
Frozen Image Encoder $f_{\theta_{orig}}$
Text
Frozen Text Encoder f_{ϕ}
Cosine Similarity
t1
t2
t3
t4
$f_{\theta_{orig}}(x)$
Alignment Space
Distribution Consistency Calibration
Cross-Modal Semantic Structure Preservation
1 2 3 4
Temperature-Scaled Distribution p_{rob}
Clean Image x
Text-Guided Adversarial Training
ϵ-
Adversarial Attack (PGD)
Adversarial Example x_{adv}
f_{θ}
Trainable Image Encoder
Text
Text Prompts t
A photo of dog
A photo of cat
.....
Frozen Text Encoder f_{ϕ}
Cosine Similarity
t1 t2 t3 t4
Robust Output Distribution on x_{adv}
Loss Optimization
Gradient Update
Alignment-Guided Loss
$\min_{\theta} \mathbb{E}_{x \in \mathcal{D}} \left[\max_{x_{adv} \in B(x,\epsilon)} L(x_{adv}, t, p_{rob}, \tau) \right]$