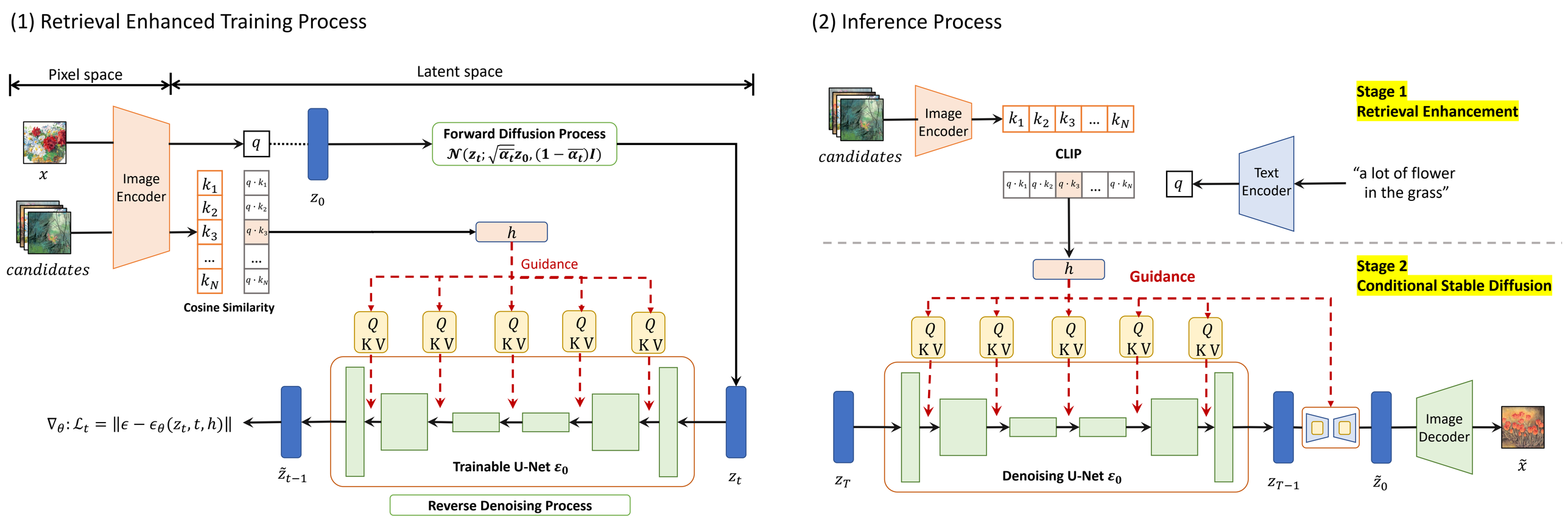
(1) Retrieval Enhanced Training Process
Pixel space
Latent space
x
Image Encoder
q
z_0
Forward Diffusion Process
$\mathcal{N}(z_t; \sqrt{\overline{\alpha}_t} z_0, (1-\overline{\alpha}_t)I)$
candidates
k_1
k_2
k_3
...
k_N
$q \cdot k_1$
$q \cdot k_2$
$q \cdot k_3$
...
$q \cdot k_N$
Cosine Similarity
h
Guidance
Q
K V
$\nabla_\theta: \mathcal{L}_t = \|\epsilon - \epsilon_\theta(z_t, t, h)\|$
$\tilde{z}_{t-1}$
Trainable U-Net ε_0
z_t
Reverse Denoising Process
(2) Inference Process
candidates
Image Encoder
k_1 k_2 k_3 ... k_N
CLIP
$q \cdot k_1$ $q \cdot k_2$ $q \cdot k_3$ ... $q \cdot k_N$
q
Text Encoder
“a lot of flower in the grass”
Stage 1
Retrieval Enhancement
h
Guidance
Stage 2
Conditional Stable Diffusion
Q
K V
z_T
Denoising U-Net ε_0
z_{T-1}
$\tilde{z}_0$
Image Decoder
$\tilde{x}$