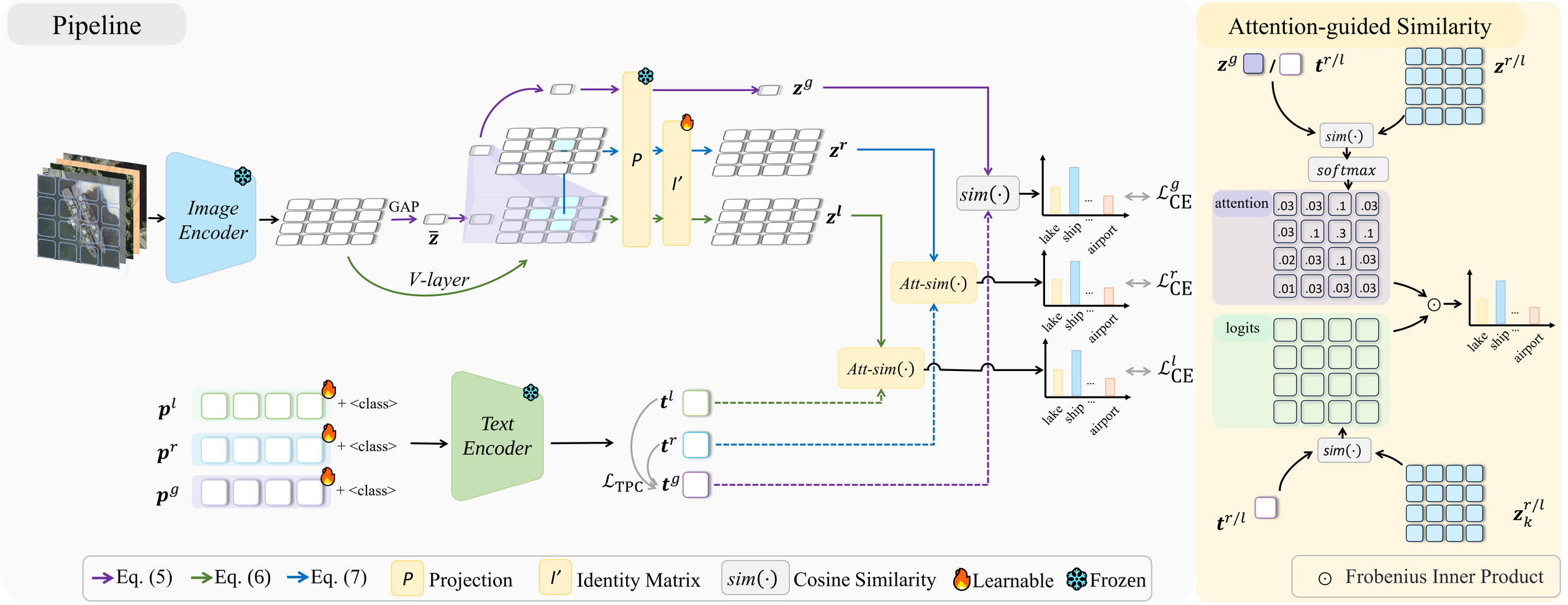
Pipeline
Image Encoder
GAP
$\bar{z}$
V-layer
P
I′
z^g
z^r
z^l
sim(·)
Att-sim(·)
Att-sim(·)
lake
ship
airport
$\mathcal{L}^g_{\text{CE}}$
$\mathcal{L}^r_{\text{CE}}$
$\mathcal{L}^l_{\text{CE}}$
p^l
p^r
p^g
+ <class>
Text Encoder
t^l
t^r
t^g
$\mathcal{L}_{\text{TPC}}$
Attention-guided Similarity
z^g / $t^{r/l}$
$z^{r/l}$
sim(·)
softmax
attention
.03 .03 .1 .03
.03 .1 .3 .1
.02 .03 .1 .03
.01 .03 .03 .03
logits
$t^{r/l}$
$z_k^{r/l}$
Eq. (5)
Eq. (6)
Eq. (7)
P Projection
I′ Identity Matrix
sim(·) Cosine Similarity
Learnable
Frozen
⊙ Frobenius Inner Product