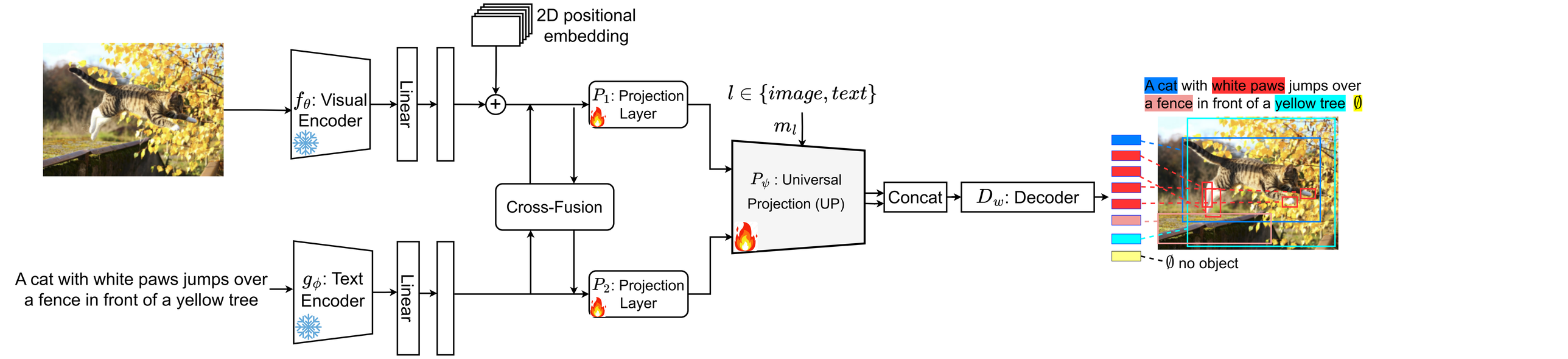
2D positional embedding
f_θ: Visual Encoder
Linear
P_1: Projection Layer
$l \in \{image, text\}$
m_l
Cross-Fusion
P_ψ : Universal Projection (UP)
Concat
D_w: Decoder
A cat with white paws jumps over a fence in front of a yellow tree ∅
A cat with white paws jumps over a fence in front of a yellow tree
g_ϕ: Text Encoder
Linear
P_2: Projection Layer
∅ no object