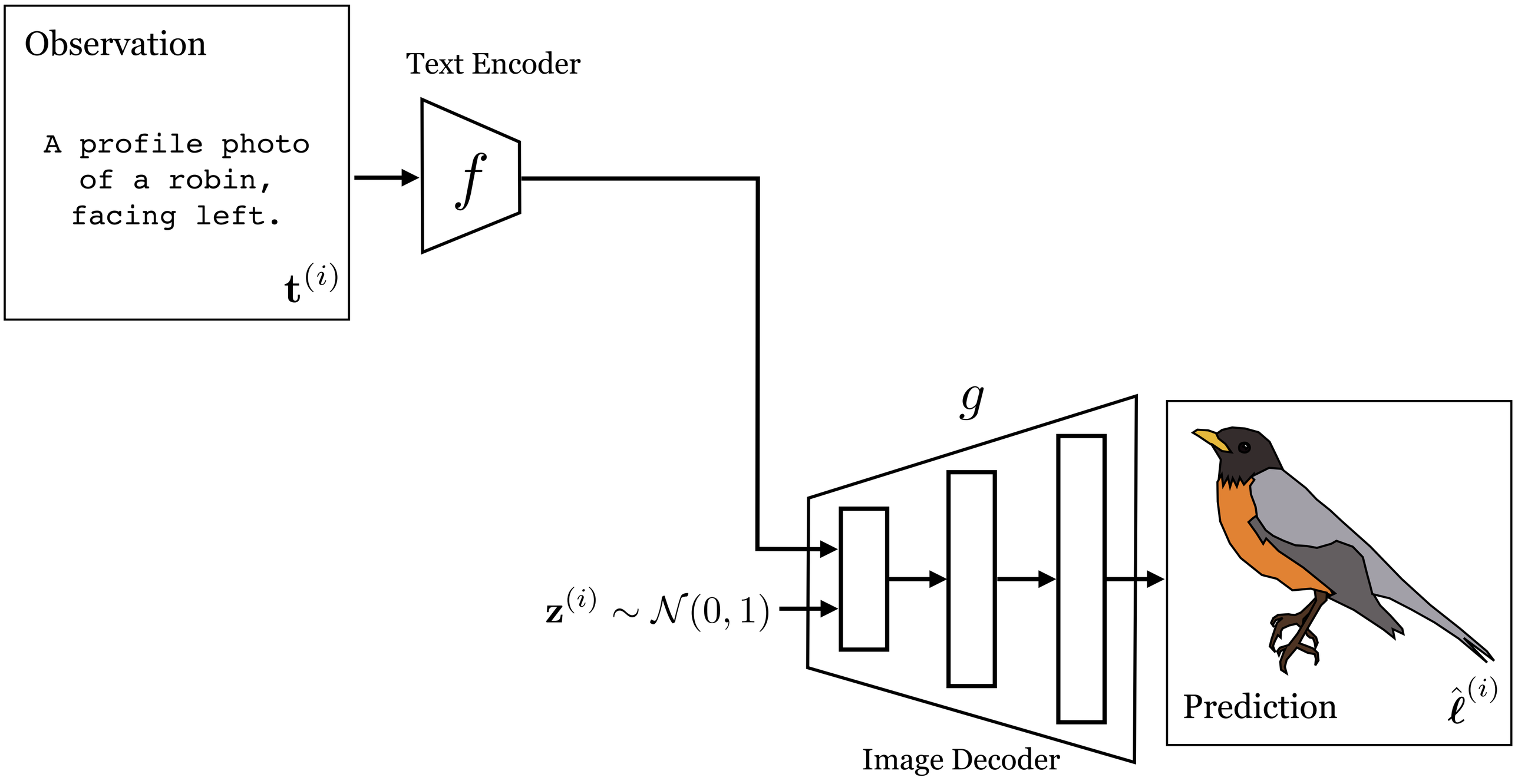
Observation
A profile photo of a robin, facing left.
$\mathbf{t}^{(i)}$
Text Encoder
f
g
$\mathbf{z}^{(i)} \sim \mathcal{N}(0, 1)$
Image Decoder
Prediction
$\hat{\boldsymbol{\ell}}^{(i)}$