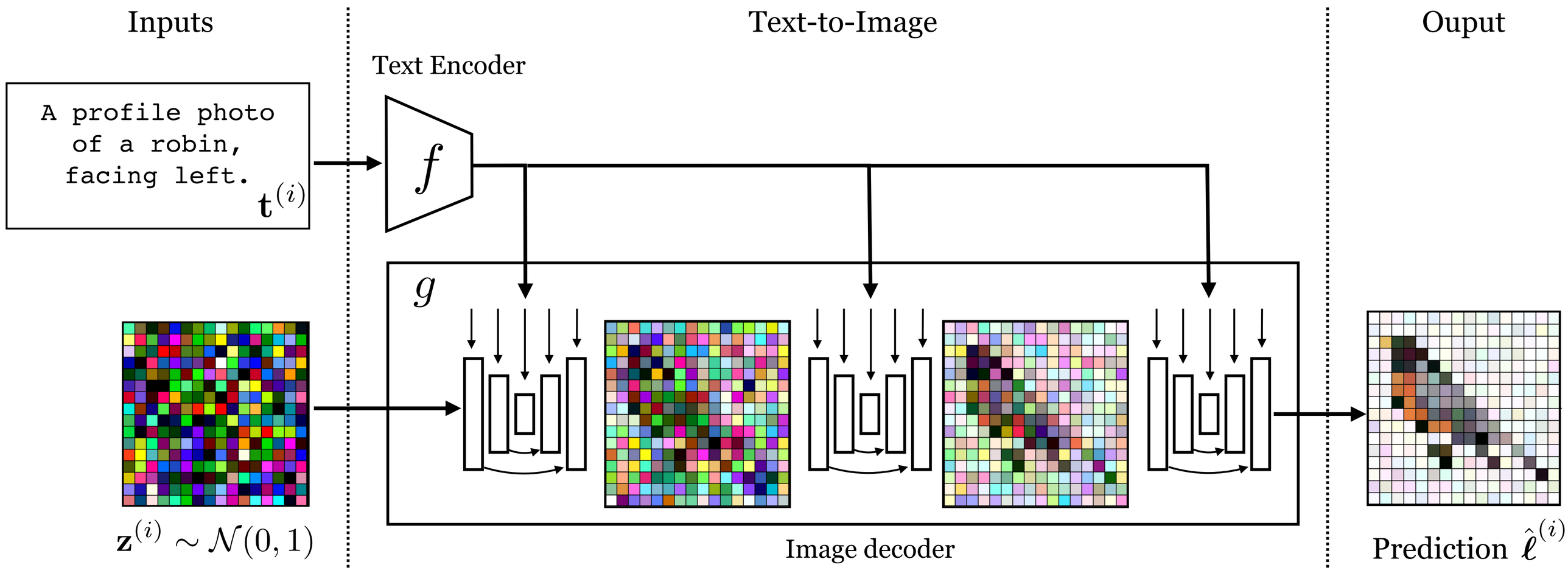
Inputs
Text-to-Image
Ouput
Text Encoder
A profile photo of a robin, facing left.
$\mathbf{t}^{(i)}$
f
g
$\mathbf{z}^{(i)} \sim \mathcal{N}(0, 1)$
Image decoder
Prediction $\hat{\ell}^{(i)}$