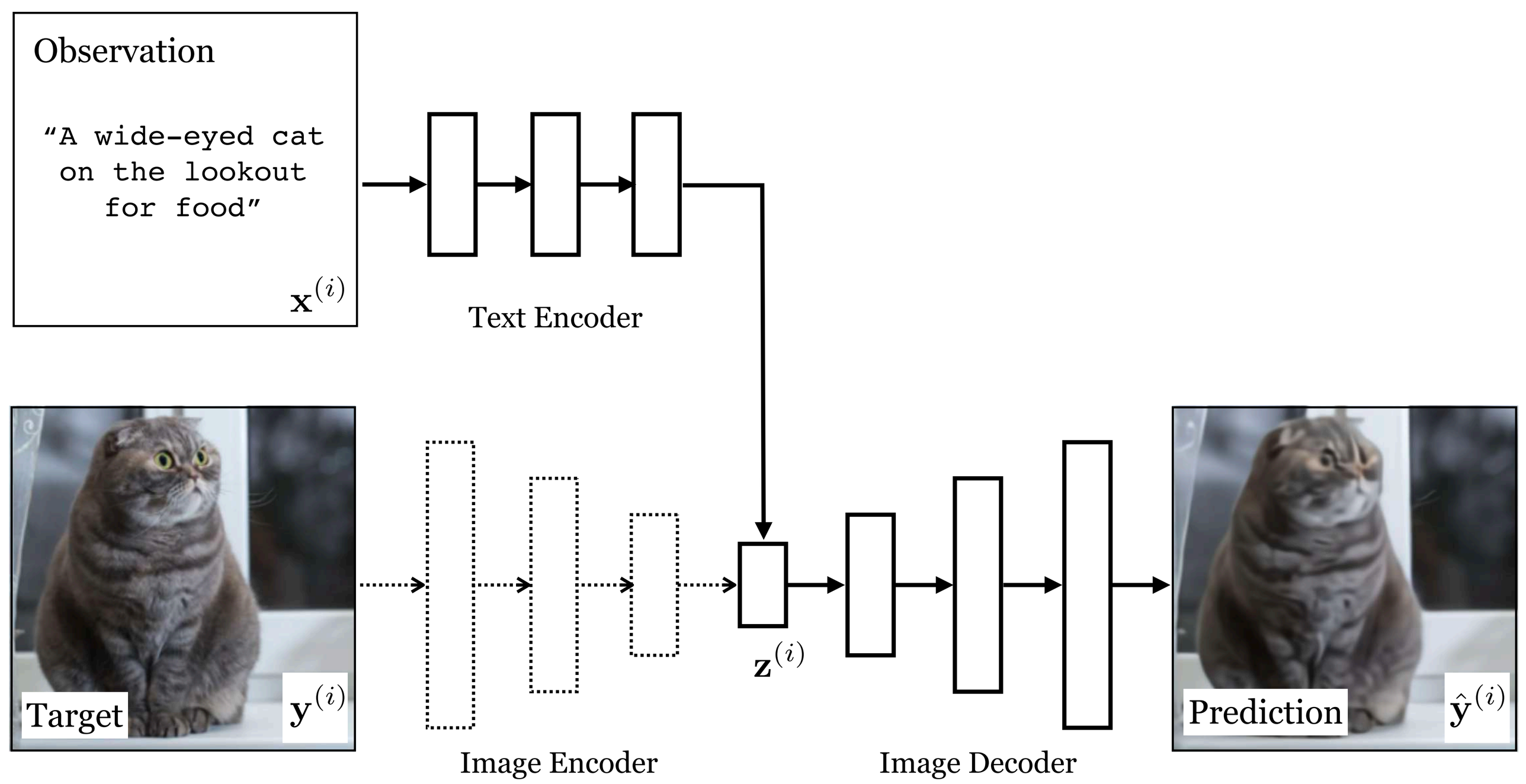
Observation
"A wide-eyed cat on the lookout for food"
$\mathbf{x}^{(i)}$
Text Encoder
Target
$\mathbf{y}^{(i)}$
$\mathbf{z}^{(i)}$
Image Encoder
Image Decoder
Prediction
$\hat{\mathbf{y}}^{(i)}$