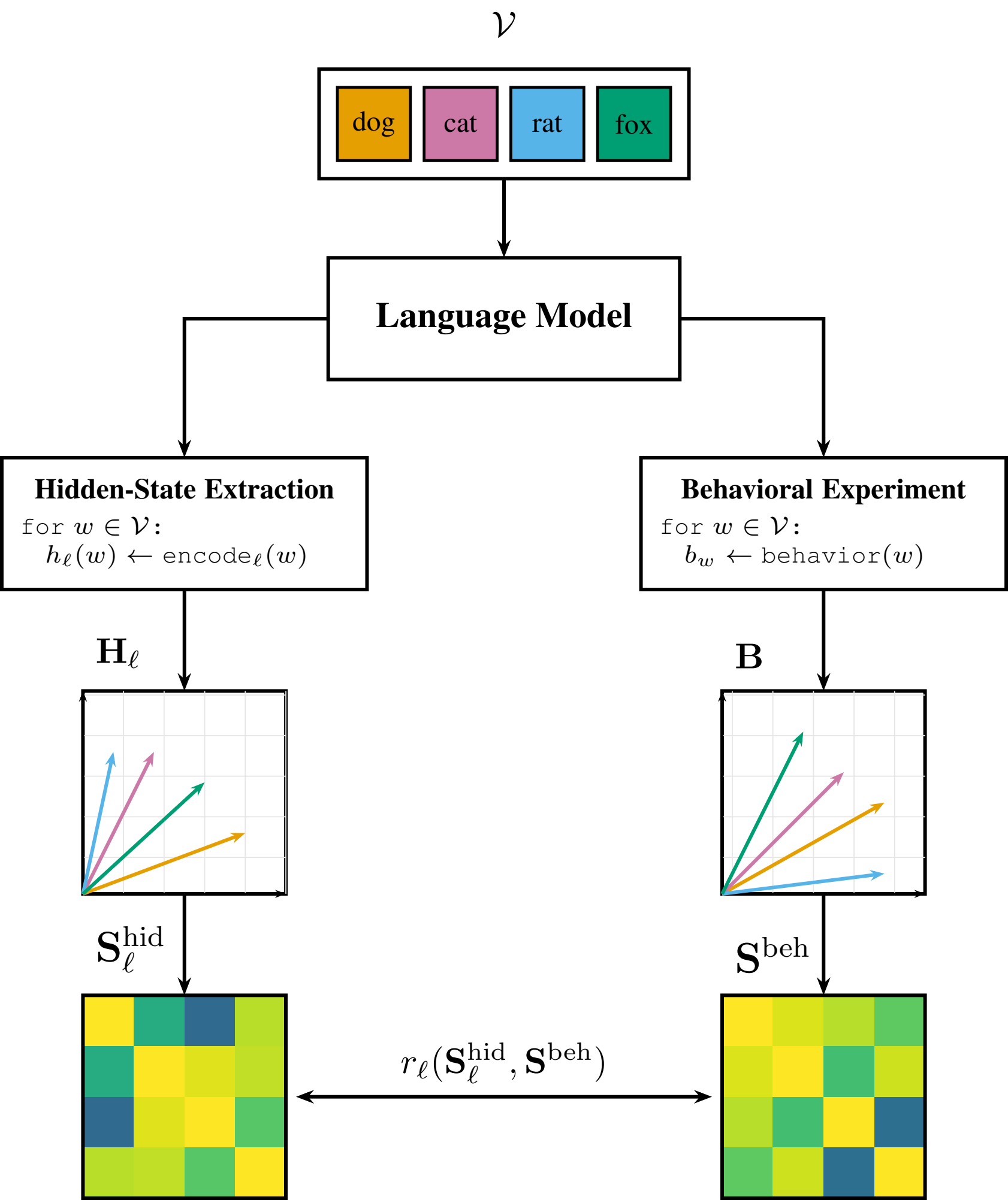
$\mathcal{V}$
dog
cat
rat
fox
Language Model
Hidden-State Extraction
for $w \in \mathcal{V}$:
$h_\ell(w) \leftarrow \texttt{encode}_\ell(w)$
Behavioral Experiment
for $w \in \mathcal{V}$:
$b_w \leftarrow \texttt{behavior}(w)$
$\mathbf{H}_\ell$
$\mathbf{B}$
$\mathbf{S}_\ell^{\text{hid}}$
$\mathbf{S}^{\text{beh}}$
$r_\ell(\mathbf{S}_\ell^{\text{hid}}, \mathbf{S}^{\text{beh}})$