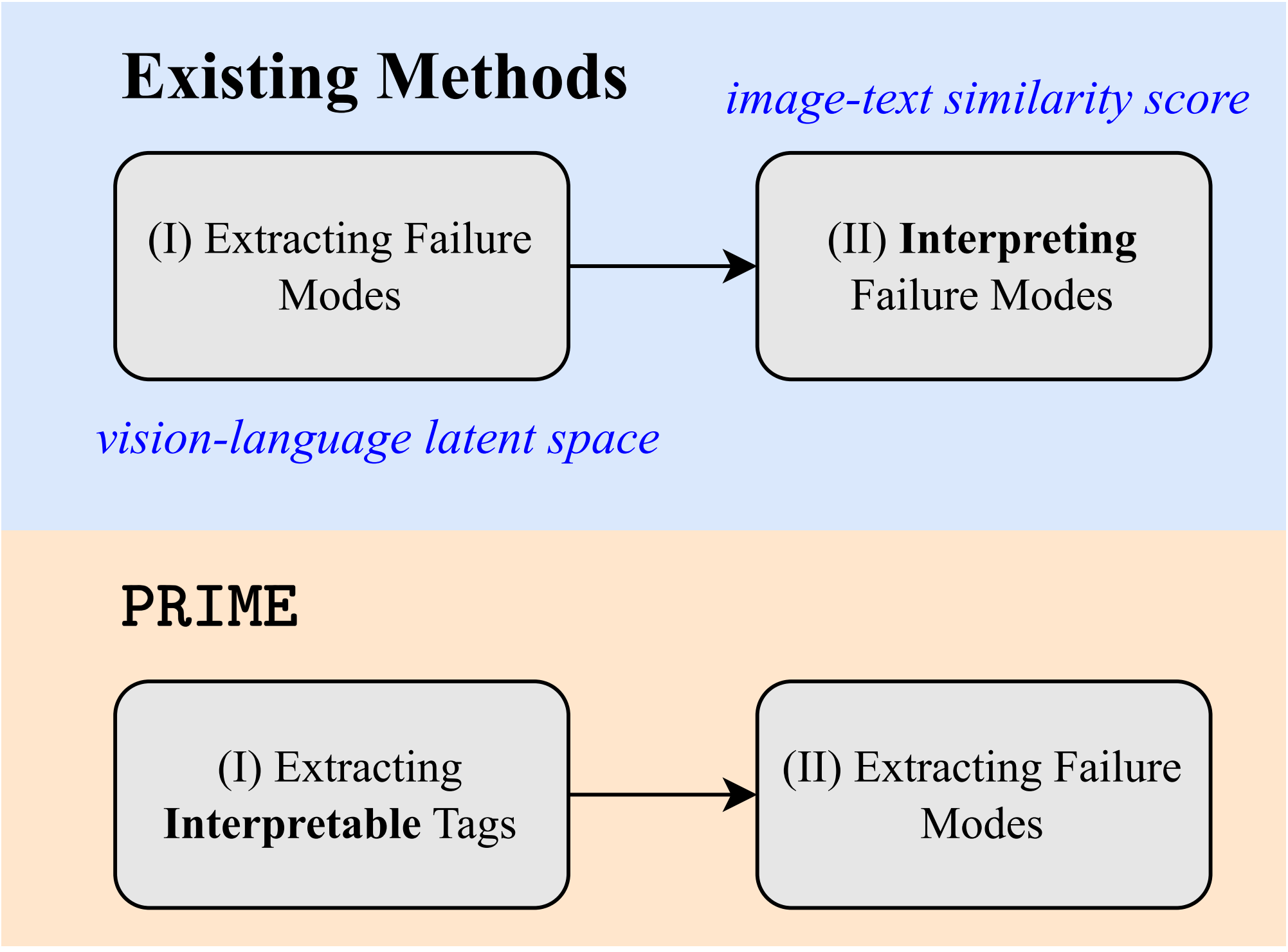
Existing Methods
image-text similarity score
(I) Extracting Failure Modes
(II) Interpreting Failure Modes
vision-language latent space
PRIME
(I) Extracting Interpretable Tags
(II) Extracting Failure Modes