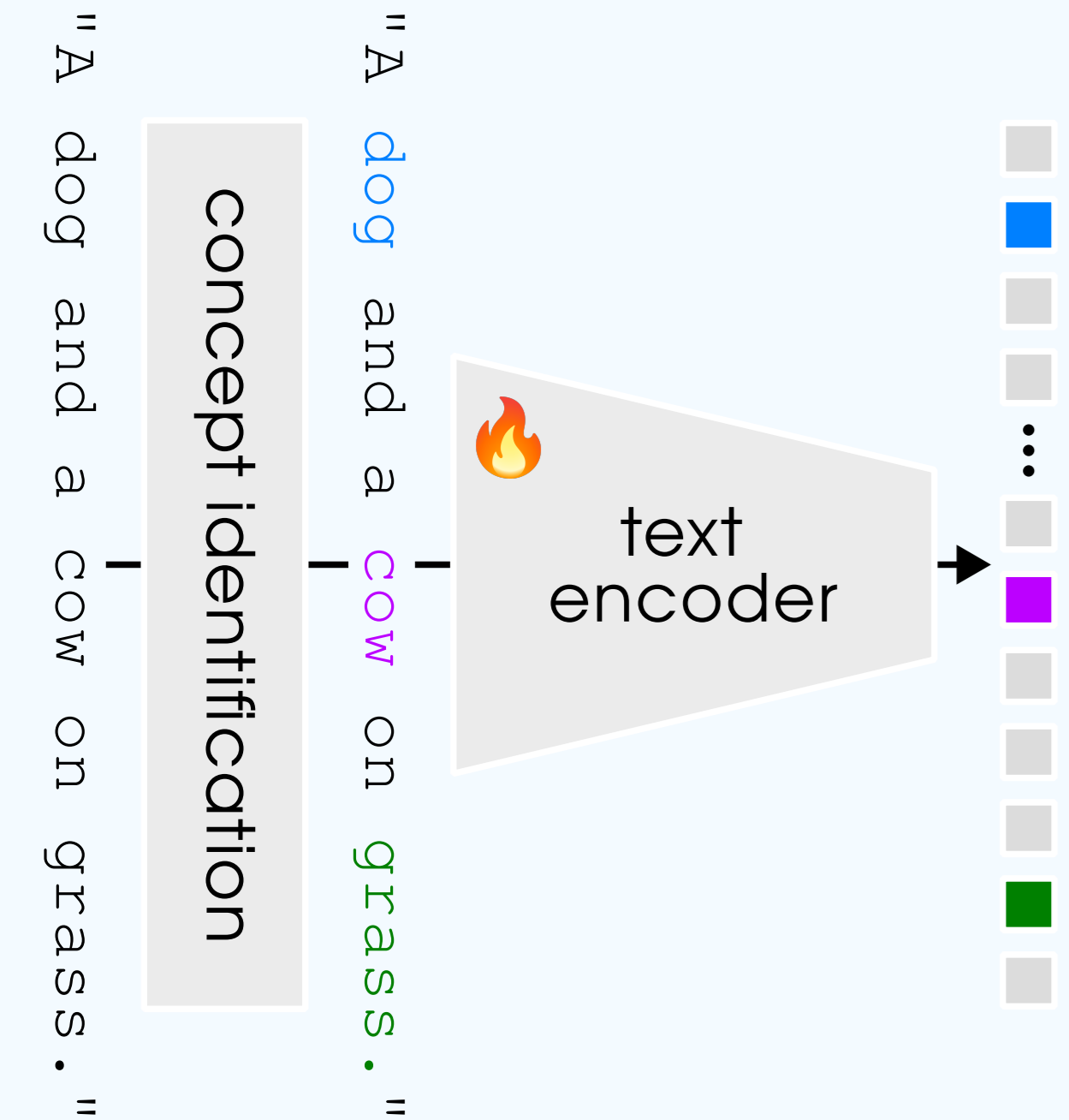
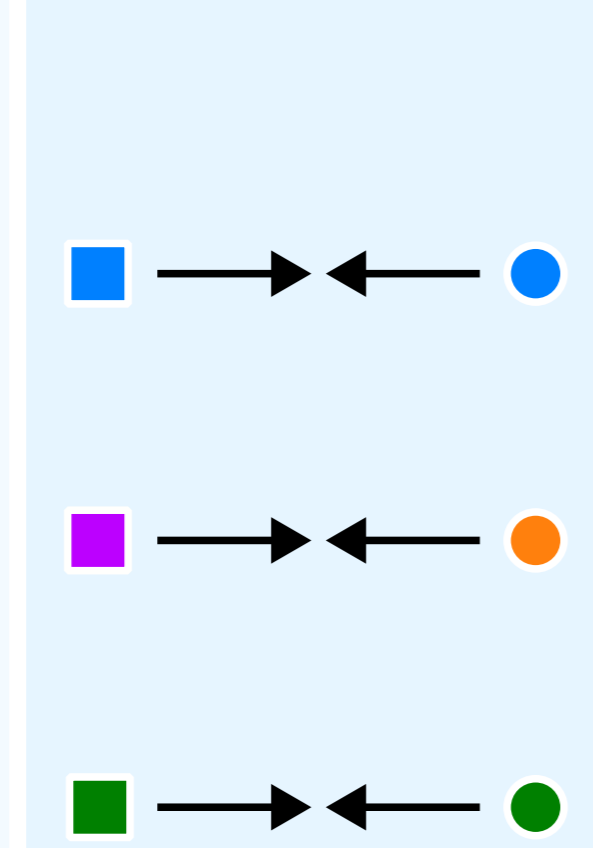


text concept extraction



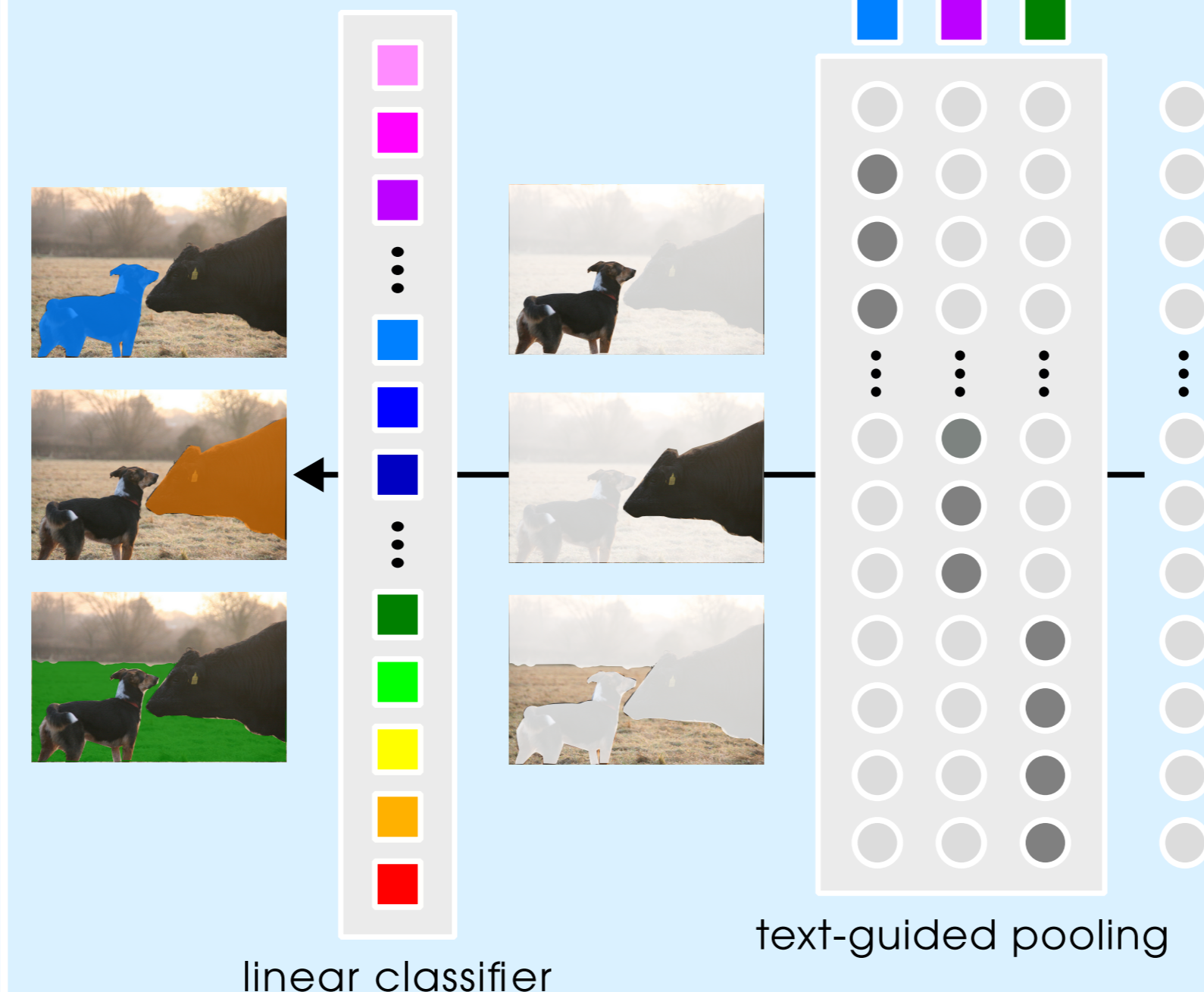
a.

cross-modality consistency

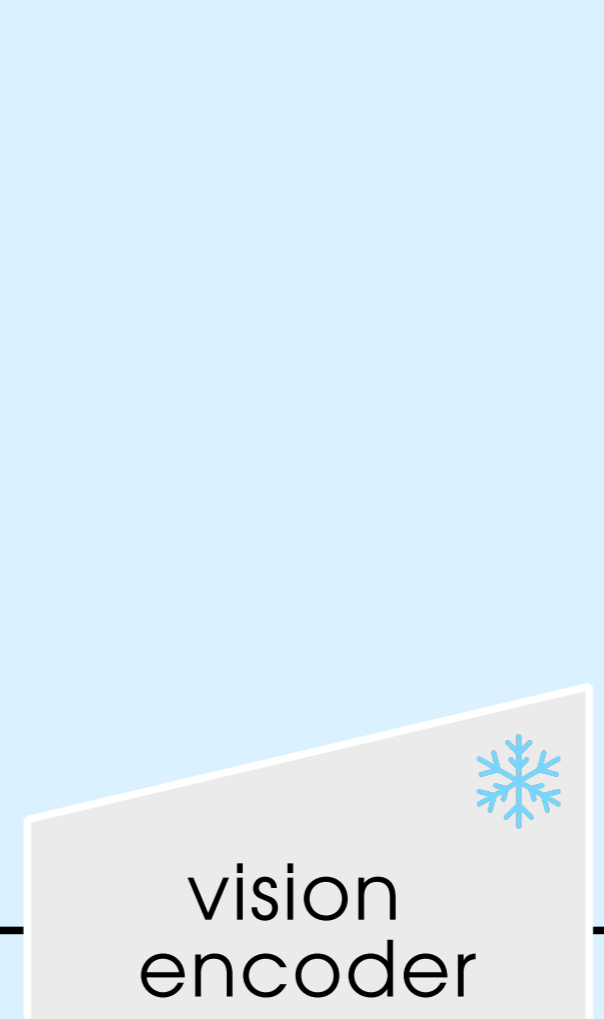


b.

vision concept extraction



c.



■ text representation
○ vision representation