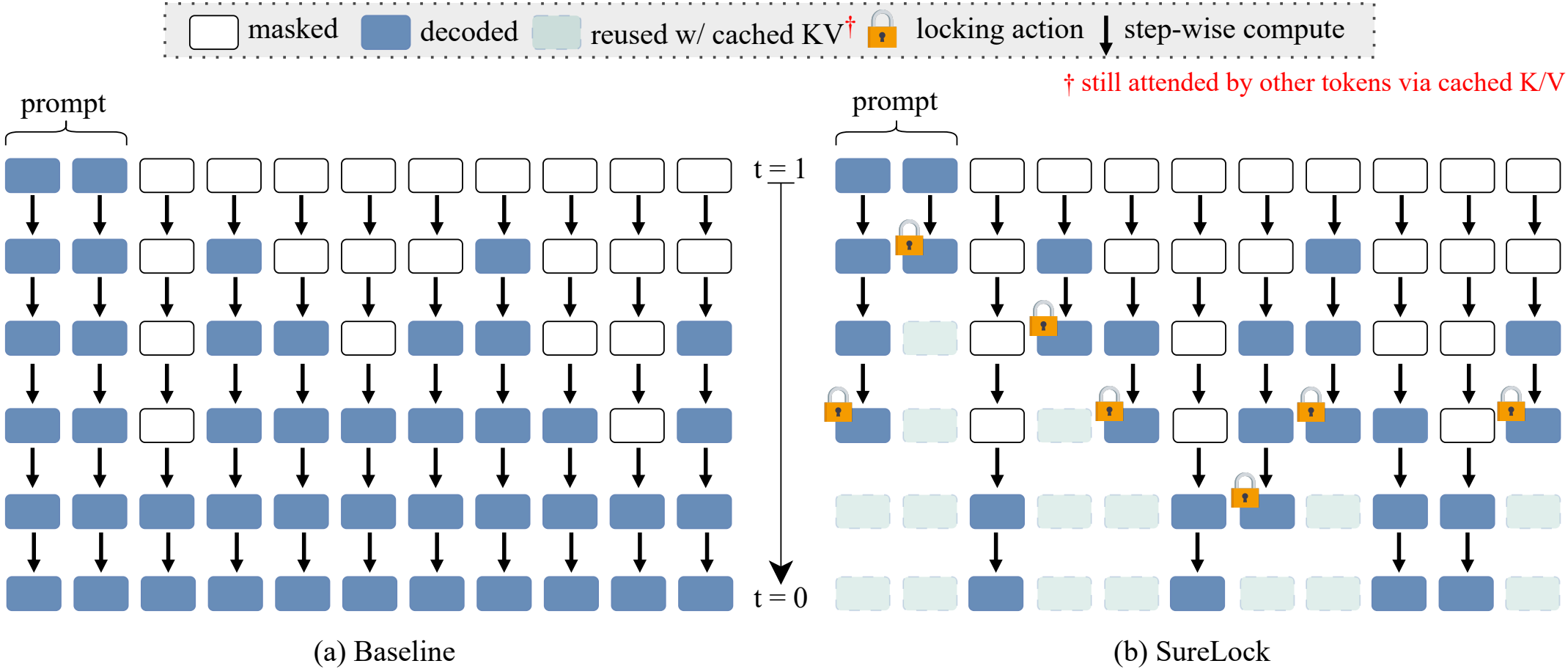
masked
decoded
reused w/ cached KV†
locking action
step-wise compute
† still attended by other tokens via cached K/V
prompt
t = 1
t = 0
prompt
(a) Baseline
(b) SureLock