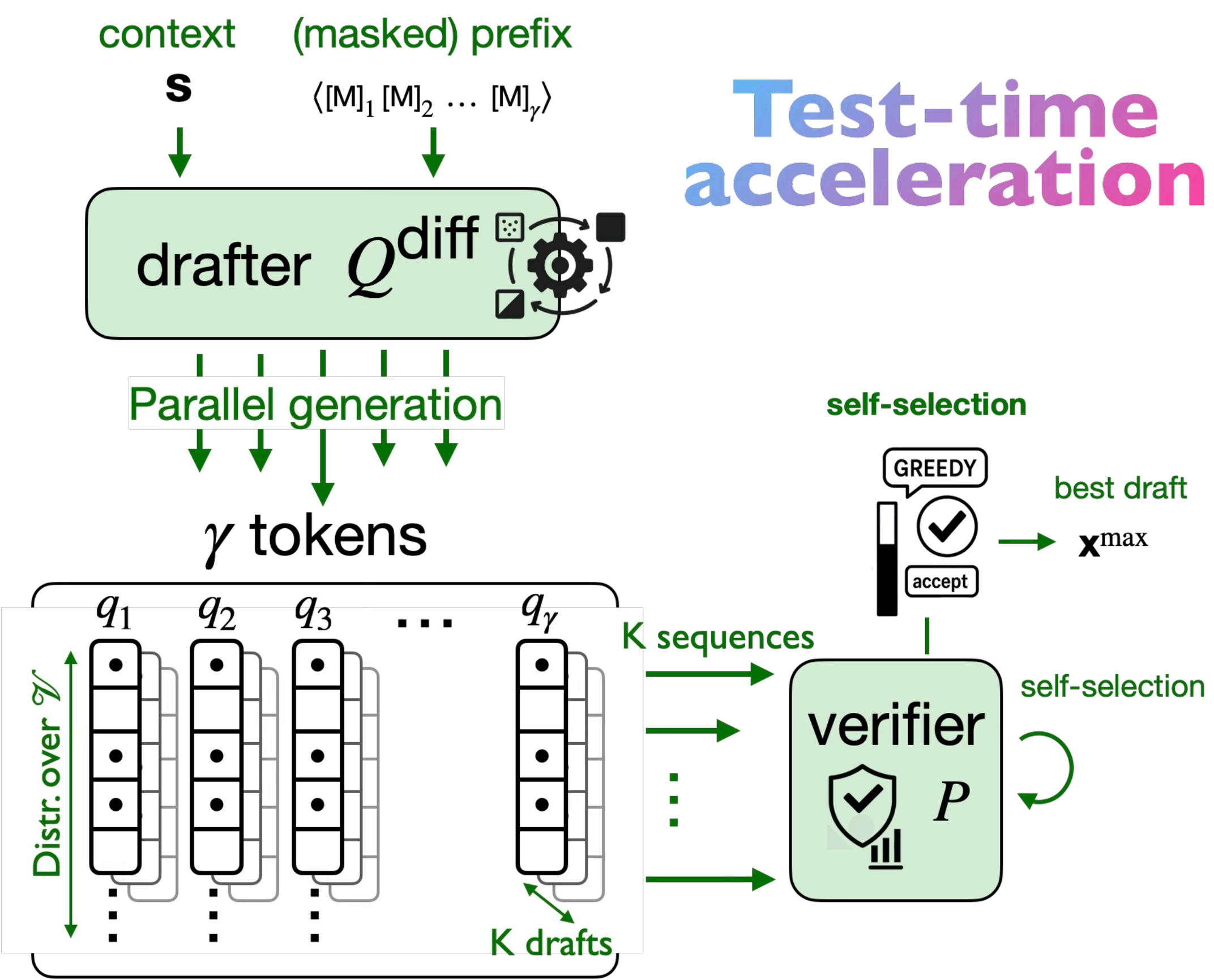
context
(masked) prefix
s
⟨[M]1 [M]2 … [M]γ⟩
Test-time acceleration
drafter Q^{diff}
Parallel generation
self-selection
GREEDY
accept
best draft
$\mathbf{x}^{\max}$
γ tokens
q_1 q_2 q_3 … q_γ
K sequences
Distr. over $\mathscr{V}$
verifier
P
self-selection
K drafts