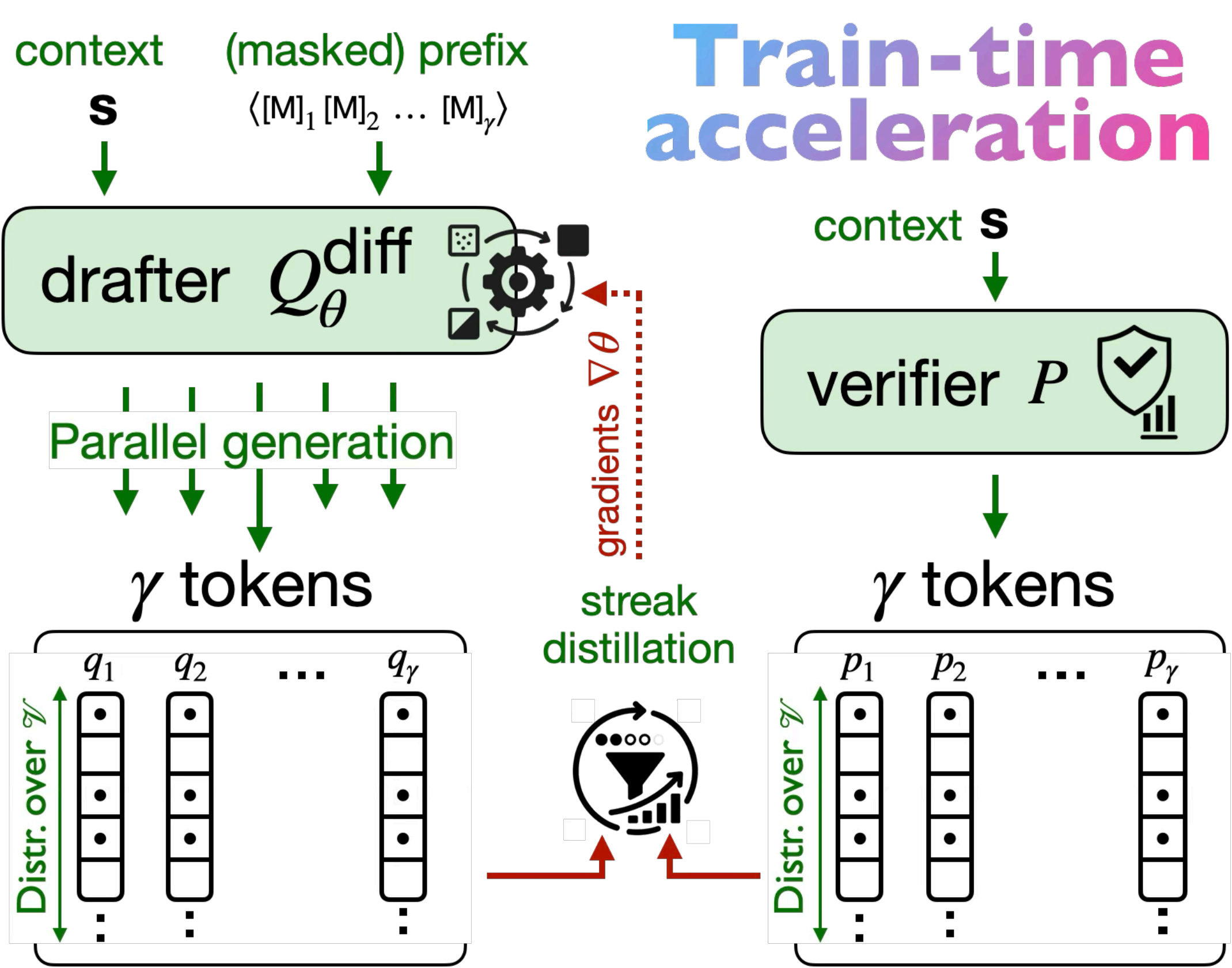
Train-time acceleration
context
(masked) prefix
s
⟨[M]1 [M]2 … [M]γ⟩
drafter Q_θ^{diff}
context s
verifier P
Parallel generation
gradients $\nabla\theta$
γ tokens
γ tokens
streak distillation
q_1 q_2 … q_γ
p_1 p_2 … p_γ
Distr. over $\mathscr{V}$
Distr. over $\mathscr{V}$