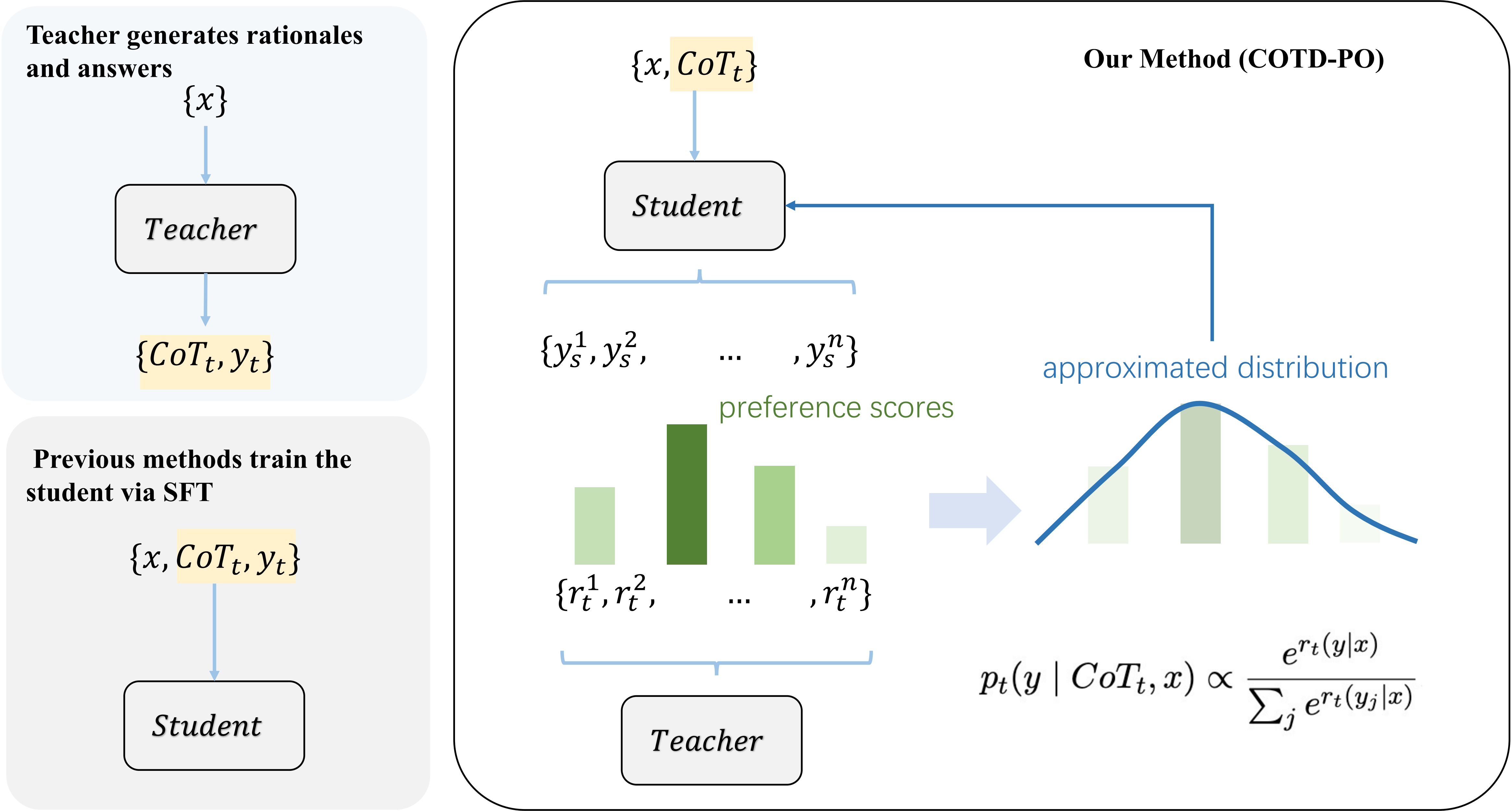
Teacher generates rationales and answers
$\{x\}$
Teacher
$\{CoT_t, y_t\}$
Previous methods train the student via SFT
$\{x, CoT_t, y_t\}$
Student
Our Method (COTD-PO)
$\{x, CoT_t\}$
Student
$\{y_s^1, y_s^2, \dots, y_s^n\}$
approximated distribution
preference scores
$\{r_t^1, r_t^2, \dots, r_t^n\}$
$p_t(y \mid CoT_t, x) \propto \frac{e^{r_t(y|x)}}{\sum_j e^{r_t(y_j|x)}}$
Teacher