

Vision: v

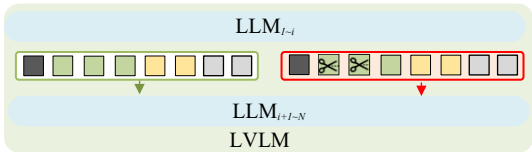
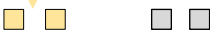


Instruction: t

Please describe...

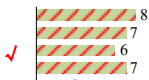
Generated: g

System: s



$\text{logit}_\theta(v, t, y_{ci})$

$\text{logit}_\theta(v_d, t, y_{ci})$

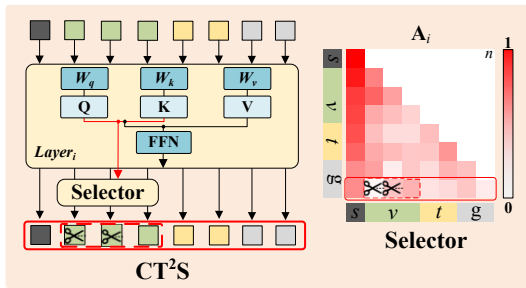


$(1+\alpha)\text{logit}_\theta(v,,)-\alpha\text{logit}_\theta(v_d,,)$

...Two persons are playing tennis

generated next

Self-Introspective Decoding (SID)



...Two persons are playing tennis



Low Score_i (Eq. 5)



High Score_i (Eq. 5)