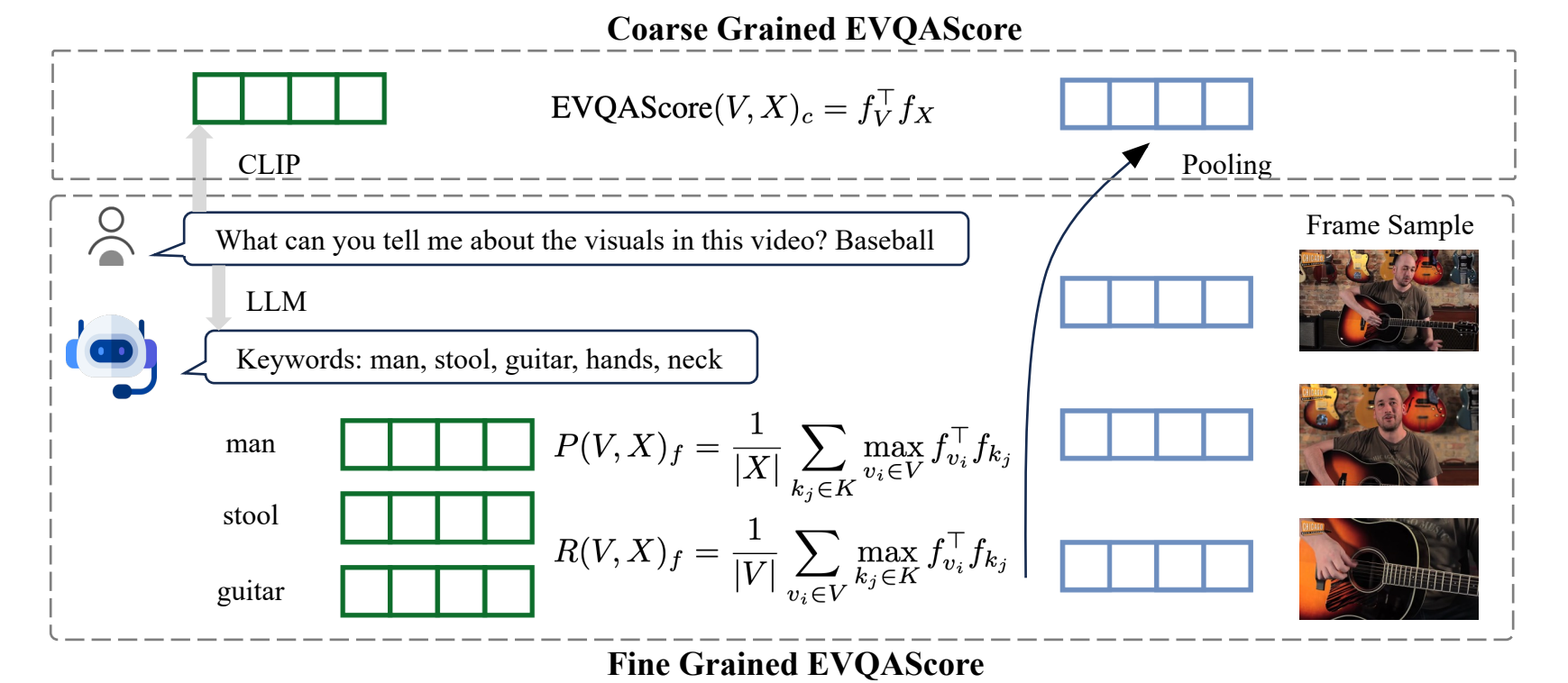
Coarse Grained EVQAScore
$\text{EVQAScore}(V, X)_c = f_V^\top f_X$
CLIP
Pooling
What can you tell me about the visuals in this video? Baseball
LLM
Keywords: man, stool, guitar, hands, neck
Frame Sample
man
stool
guitar
$P(V, X)_f = \frac{1}{|X|} \sum_{k_j \in K} \max_{v_i \in V} f_{v_i}^\top f_{k_j}$
$R(V, X)_f = \frac{1}{|V|} \sum_{v_i \in V} \max_{k_j \in K} f_{v_i}^\top f_{k_j}$
Fine Grained EVQAScore