Next-token prediction — Cross-Entropy loss

Segment prediction — L1 loss

Time Refinement Sequence

From <seg_start> 30 to 46 <offset> +05 and -06 <refine>
32 to 44 <offset> +03 and -04 <refine>
35 to 41 <offset> +00 and -01 <refine>
35 to 40 <offset> +00 and -00 <seg_end>, the person stirs soup.

Aux. Pred. Head → [start,end]

Aux. Pred. Head → [start,end]

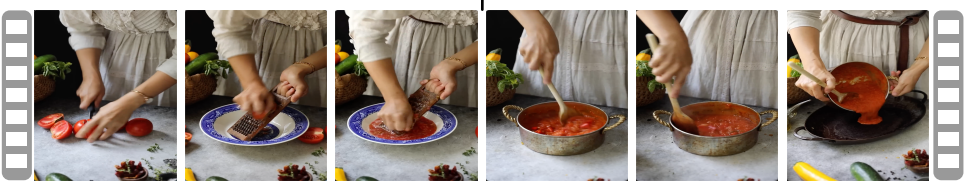
Aux. Pred. Head → [start,end]

Large Language Model

LoRA

Visual Adapter

Visual Encoder



Tokenizer

User Prompt

When does the person stir soup in the video?