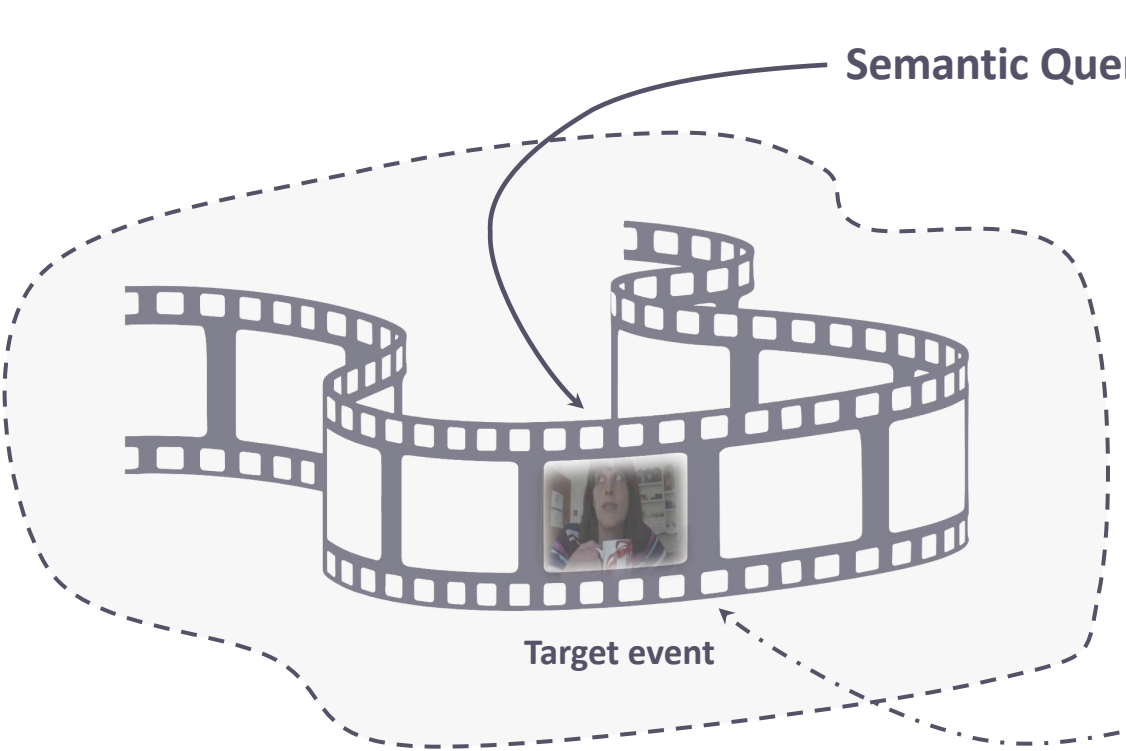

So far, we only use *Natural Language Query* q_t:

Sentence: Woman holds her a lobster coffee mug.

What comes next: *Multimodal Query* q_m:

(Reference image + *Refinement text* TXT *)*

User scribble

Similar scenes

Past

Future