Algorithm 1: Confidence-based Event-centric Online Video Question Answering.

Inputs : Online video stream $\mathcal{O}$, question $\boldsymbol{q}$, answer space $\mathcal{A}$
Output: Answer $\boldsymbol{\alpha}$ to the question

Initialize confidence thresholds c_{min}, c_{max};
for *Every frame* $\boldsymbol{f} \in \mathcal{O}$ **do**
    $c \leftarrow$ `VideoTextEncoder(`$\boldsymbol{q}, \boldsymbol{f}$`)`;
    if $c \geq c_{max}$ **then**
        $\mathcal{T}_B :=$ *Backward-traverse frames till* $c < c_{min}$;
        $\mathcal{T}_F :=$ *Forward-traverse frames till* $c < c_{min}$;
        $\mathcal{T} := \mathcal{T}_B \cup \mathcal{T}_F$;
        Break;
    end
end
Video features $\boldsymbol{V}$:=`VideoEncoder(`$\mathcal{T}$`)`;
Question features $\boldsymbol{Q}$:=`QuestionEncoder(`$\boldsymbol{q}$`)`;
$\boldsymbol{M}$:=`MultiModalEncoder(`$\boldsymbol{Q}, \boldsymbol{V}$`)`;
$\boldsymbol{\alpha} \leftarrow$`AnswerDecoder(`$\boldsymbol{M}$, $\mathcal{A}$`)`;