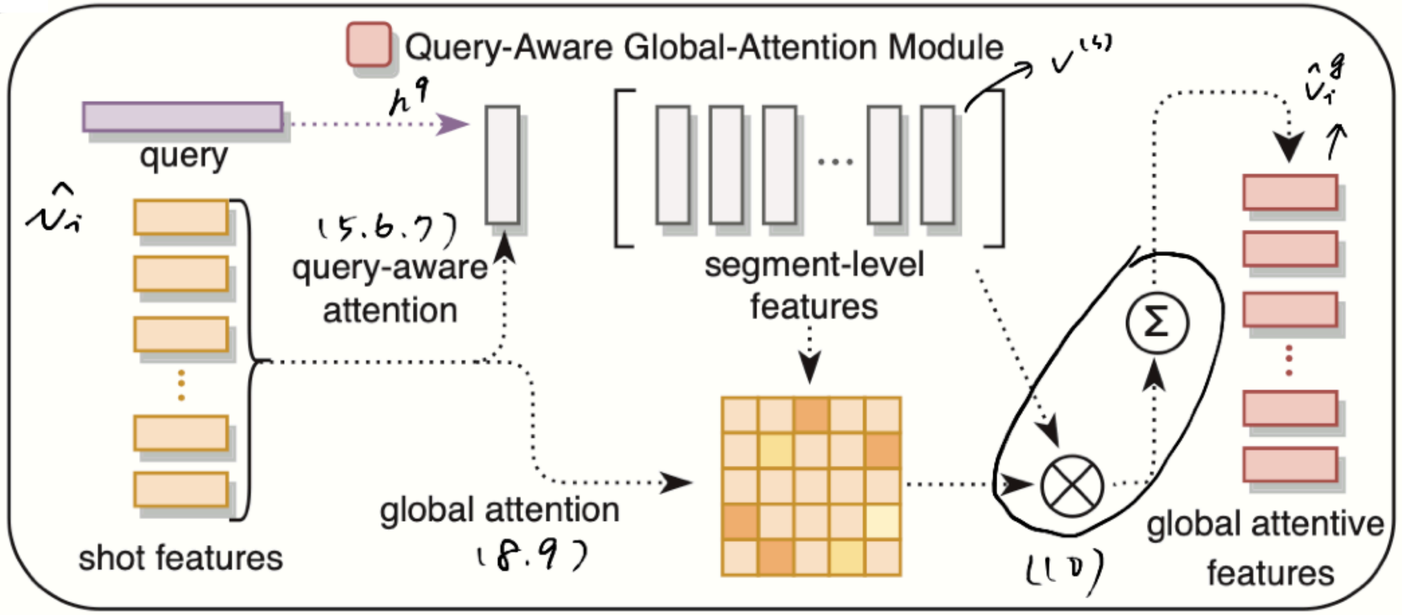
Proposed Method

Query global-attention module

- Model the relationship of different video segments among the video and to generate query-focused visual representation.
- Given $(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_t)$ and query q (composed of two concept (c_1, c_2))



(5) $e_i = v^T \tanh(W_1 \hat{v}_i + W_2 h^q + b)$

- v^T, W_1, W_2 : trainable parameters, b : bias vector
- h^q : average of representation of concepts

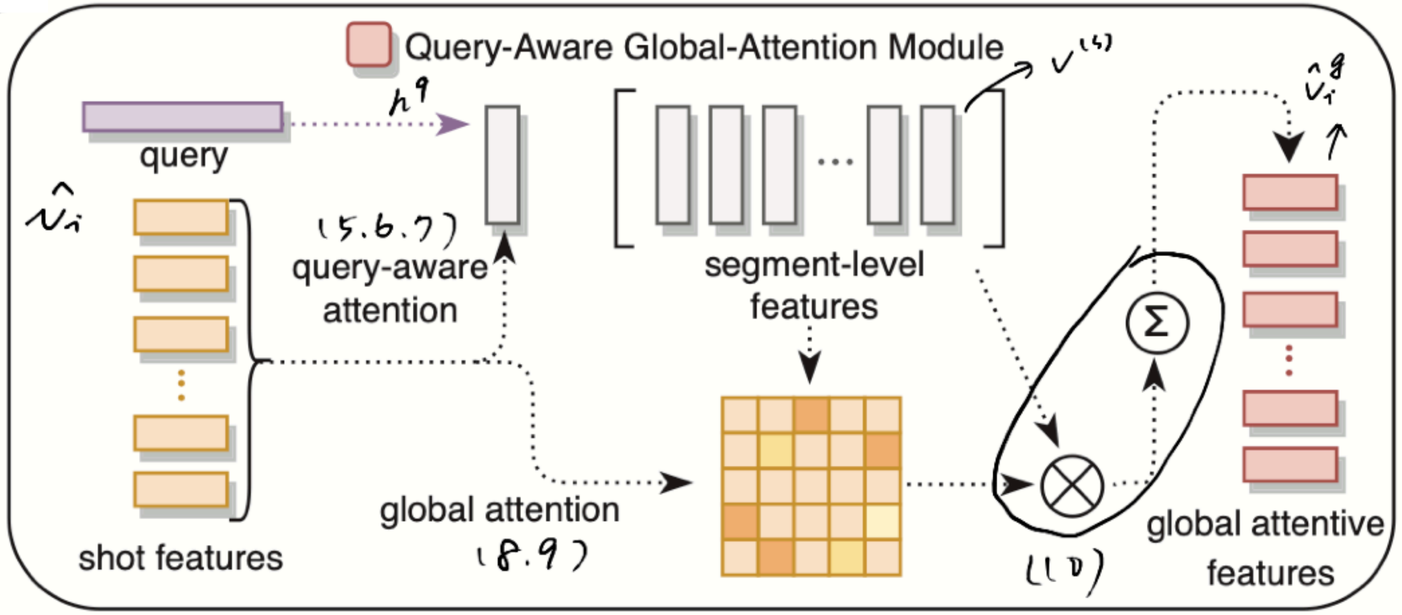
(6) $r_i = \dfrac{\exp(e_i)}{\sum_{k=0}^{t} \exp(e_k)}$

(7) Segment-level visual feature: $v^{(s)} = \sum_{i=0}^{t} r_i \hat{v}_i$

Proposed Method

Query global-attention module

- Compute the query-aware global -attentive representation for each shot.
- Given visual feature $\hat{v}_i$ & all segment-level visual representation $(v_1^{(s)}, v_2^{(s)}, \ldots, v_m^{(s)})$
  - m: number of video segments



(8) $e_j^g = v^T \tanh(W_1^g \hat{v}_i + W_2^g v_j^{(s)} + b)$

(9) $r_j^g = \dfrac{\exp(e_j^g)}{\sum_{k=0}^{m} \exp(e_k^g)}$

(10) Global-attentive representation for i-shot:

$$\hat{v}_j^g = \sum_{j=0}^{m} r_j^g v_j^s$$