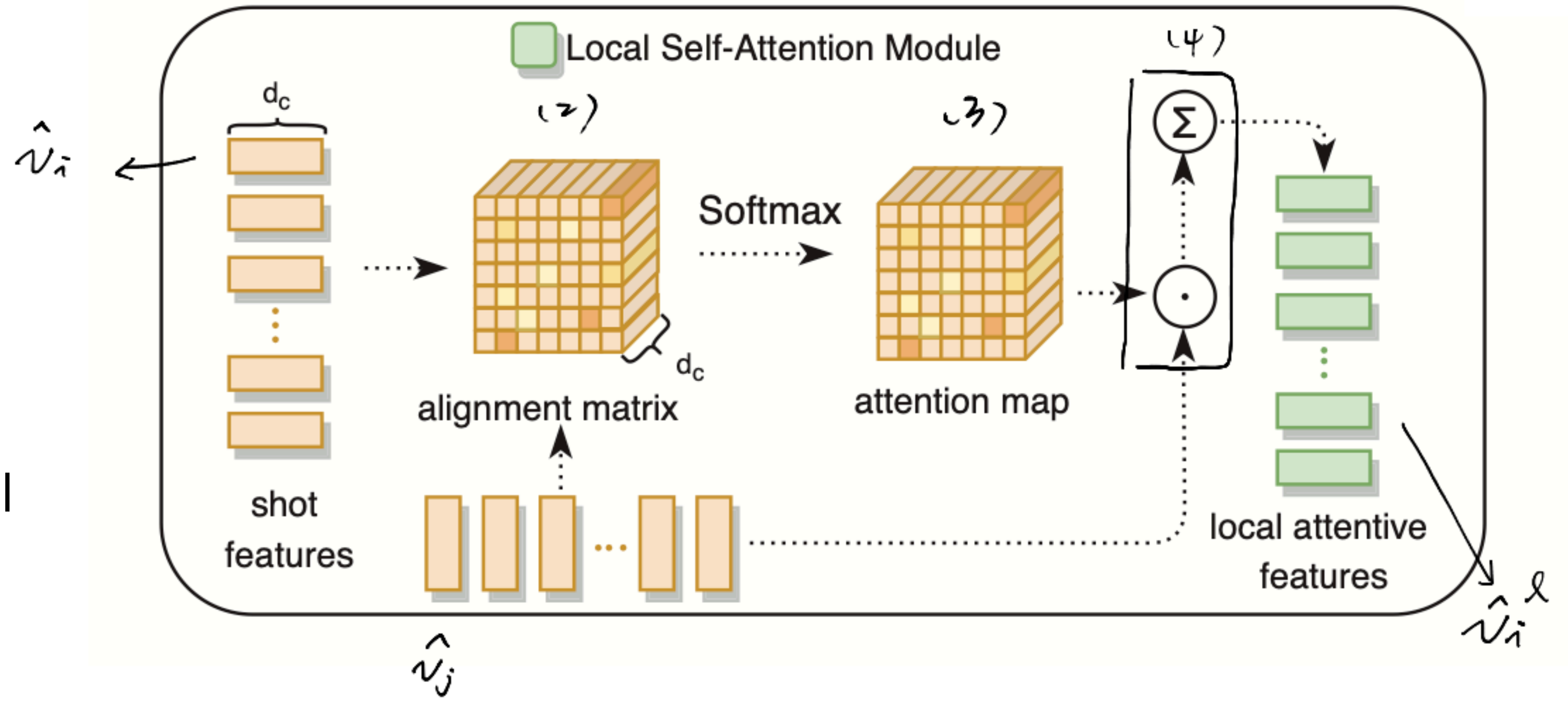
Proposed Method

Local self-attention module



- Capture the semantic relations between all shots among a video segment.
- Given $(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_t)$ to compute the alignment matrix. (shape: $t \times t \times d_c$)
- Module can learn the relative semantic relationship of different frames in the same segments.
- For different segments, the relation structure should be similar. Therefore, modules share all the trainable parameters, also reduces the amounts of parameters in our model.

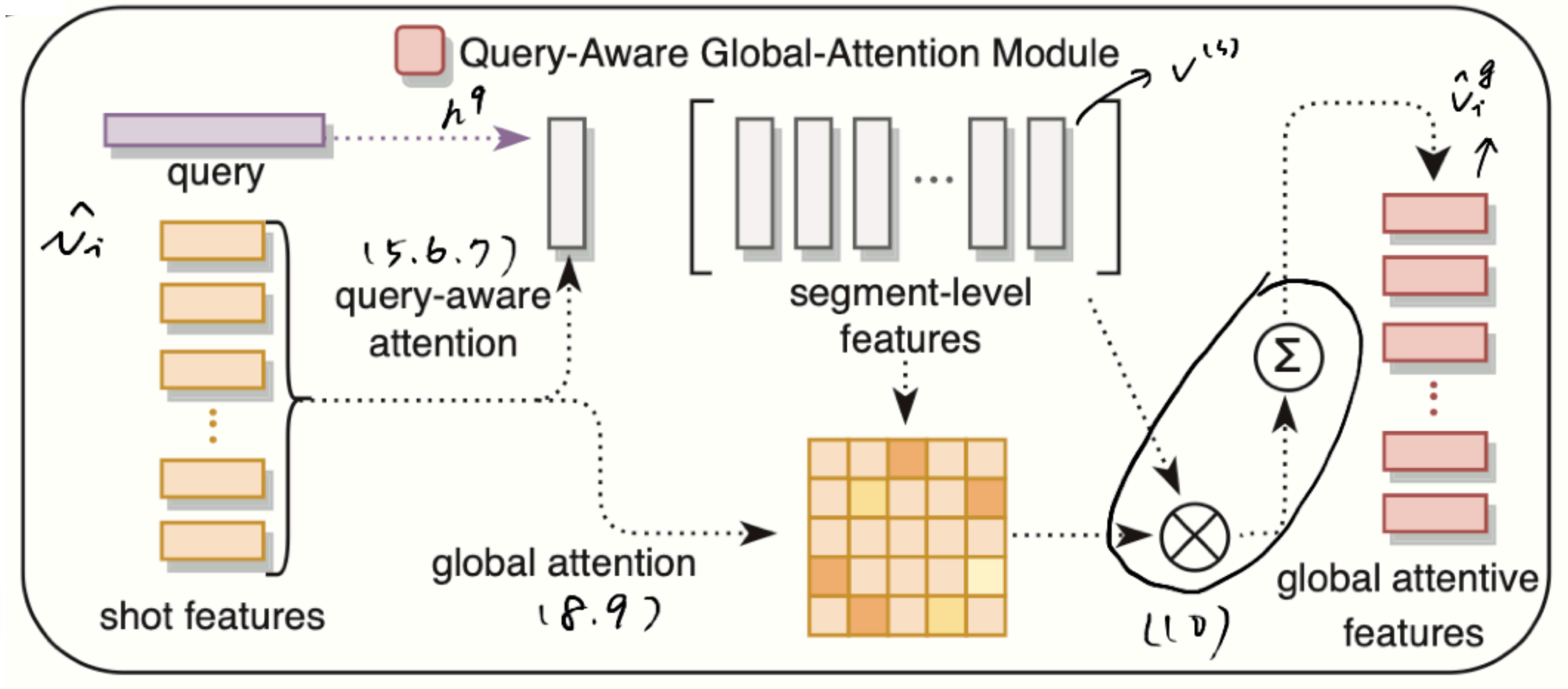
(2) $f(\hat{v}_i, \hat{v}_j) = P \tanh(W_1\hat{v}_i + W_2\hat{v}_j + b) \in R^{d_c}$

- $P, W_1, W_2 \in R^{d_c \times d_c}$: trainable parameters
- $b \in R^{d_c}$: bias vector , d_c : dimension of $\hat{v}_i$

(3) $r_{ij} = \dfrac{\exp(f(\hat{v}_i, \hat{v}_j))}{\sum_{k=0}^{t} \exp(f(\hat{v}_i, \hat{v}_k))}$

(4) Local attentive video feature for i-th: $\hat{v}_i^l = \sum_{j=0}^{t} r_{ij} \odot \hat{v}_j$

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: $v^{(s)} = \sum_{i=0}^{t} r_i \hat{v}_i$