A Spatio-Temporal Character Filter

To remove mispredictions and duplicate detections in the character recognition module, we apply a spatio-temporal filter by finding all the faces belonging to the same person and re-assigning character names based on temporal coherence. For each scene, we first detect frames belonging to the same shot (i.e., captured with the same camera) by computing the cosine similarity between their ResNet50 representations, $\mathbf{f}_{\text{res}} \in \mathbb{R}^{2048}$. Then, for all the frames in a shot, we compute the Euclidean distance between all possible pairs of faces using their bounding box centroids. When a pair of faces is close (i.e., the distance is below a threshold[9]), faces are considered to be the same. For all the faces assigned to the same person, we find the most frequent name predicted by the kNN classifier. If the name appears in at least 70% of the frames in the shot, the character is assigned to that name. Otherwise, the character is assigned as *unknown*.

B Fusion Methods Details

Given the output scores for each branch, ${\alpha_r}^c$, ${\alpha_o}^c$, and ${\alpha_{ll}}^c$, the final score for each candidate answer ω^c is computed according to each fusion method as:

- *Average*: $\omega^c = \frac{1}{3}({\alpha_r}^c + {\alpha_o}^c + {\alpha_{ll}}^c)$
- *Maximum*: $\omega^c = \max({\alpha_r}^c, {\alpha_o}^c, {\alpha_{ll}}^c)$
- *Self-att*: According to the Transformers embeddings ${\mathbf{y}_r}^c$, ${\mathbf{y}_o}^c$, and ${\mathbf{y}_{ll}}^c$, with ${\mathbf{y}_{ll}}^c = \mathbf{y}_{ll}{}^c_{j_{\max}}$ and $j_{\max} = \arg\max_j(\alpha_{ll}{}^c_j)$, the self-attention weights are computed as:
$$\boldsymbol{\psi}^c = \mathbf{W}_{\text{self}} \begin{pmatrix} {\mathbf{y}_r}^c \\ {\mathbf{y}_o}^c \\ {\mathbf{y}_{ll}}^c \end{pmatrix} + \mathbf{b}_{\text{self}}$$
with $\boldsymbol{\psi}^c = [{\psi_r}^c, {\psi_o}^c, {\psi_{ll}}^c]$. The information from the three branches is fussed according to the self-attention weights as $\mathbf{y}^c_{\text{self}} = {\psi_r}^c {\mathbf{y}_r}^c + {\psi_o}^c {\mathbf{y}_o}^c + {\psi_{ll}}^c {\mathbf{y}_{ll}}^c$, and the final score for each candidate answer is $\omega^c = {\mathbf{w}_c}^\top \cdot {\mathbf{y}_{\text{self}}}^c + b_c$.
- *QA-att*: According to the output of the pre-trained BERT network when using the question and a candidate answer, ${\mathbf{y}_{\text{QA}}}^c$, the attention weight for the read branch is computed as:
$${\psi_r}^c = {\mathbf{w}_{\text{att}}}^\top \cdot \begin{pmatrix} {\mathbf{y}_r}^c \\ {\mathbf{y}_{QA}}^c \end{pmatrix} + b_{\text{att}}$$
with ${\psi_o}^c$ and ${\psi_{ll}}^c$ computed equivalently. The information from the three branches is fused according to the attention weights as $\mathbf{y}^c_{\text{att}} = {\psi_r}^c {\mathbf{y}_r}^c + {\psi_o}^c {\mathbf{y}_o}^c + {\psi_{ll}}^c {\mathbf{y}_{ll}}^c$, and the final score for each candidate answer is computed as $\omega^c = {\mathbf{w}_c}^\top \cdot {\mathbf{y}_{\text{att}}}^c + b_c$.
- *Linear w/o MW*: A linear layer trained without MW mechanism fuses the three branches information, $\omega^c = {\mathbf{w}_c}^\top \cdot [{\alpha_r}^c, {\alpha_o}^c, {\alpha_{ll}}^c]^\top + b_c$ and $\beta_\omega = 1$.
- *Linear w/ MW* A linear layer trained with our MW mechanism fuses the three branches information, $\omega^c = {\mathbf{w}_c}^\top \cdot [{\alpha_r}^c, {\alpha_o}^c, {\alpha_{ll}}^c]^\top + b_c$ and $\beta_\omega < 1$.

[9] Experimentally set to 50.

C Video Modules Accuracy

We evaluate the performance of the different video modules used in the observe branch. As datasets in video story question answering do not generally contain annotations on characters, places, objects relationships or actions, accuracy results are difficult to obtain. Even so, we report accuracy for the Character Recognition (Table 6) and the Place Classification (Table 7) modules. For the Character Recognition, we randomly annotate a subset of 100 video frames and measure accuracy as the number of correct predictions before and after applying the spatio-temporal filter. For the Place Classification, we report accuracy before and after the temporal filter on the KnowIT VQA test set according to the labels obtained from the scripts. Additionally, visual results with the predictions from the four modules (Character Recognition, Place Classification, Object Relation Detection, and Action Recognition) are in Fig. 8. Finally, the list of characters and places can be found in Tables 8 and 9, respectively.

Fig. 8: Predictions of the video modules. For each frame, faces and objects on their bounding boxes, location on the top right, and action on the bottom middle.

Table 6: Character Recognition accuracy on 100 random frames.

	Precision	**Recall**	**F1-score**
Before filter	86.11	89.90	87.96
After filter	75.76	73.99	74.86

Table 7: Place Classification accuracy on the KnowIT VQA test set.

	Prec@1	**Prec@5**
Before filter	63.13	84.87
After filter	66.43	-

Table 8: List of characters.

Sheldon	Leonard	Penny	Howard	Raj
Amy	Bernadette	Dr. Beverly Hofstadter	Stuart	Barry
Emily	Leslie	Lucy	Mary Cooper	Priya
Dr. VM Koothrappali	Wil Wheaton			

Table 9: List of places.

Penny's apartment	The main building
Sheldon and Leonard's apartment	Penny's apartment door
A lab	A restaurant
A party	A car
Sheldon's bedroom	An office
The Cheesecake Factory	A room
Leonard's bedroom	Howard's bedroom
The cinema	The Caltech cafeteria
Caltech University	Sheldon's office
A store	The hospital
Raj's apartment	The laundry room
Penny's bedroom	The comic book store
A house	A bathroom
Outside a house or a building	A bar
Howard's house	Amy's apartment
Howard and Bernadette's apartment	Howard and Bernadette's house