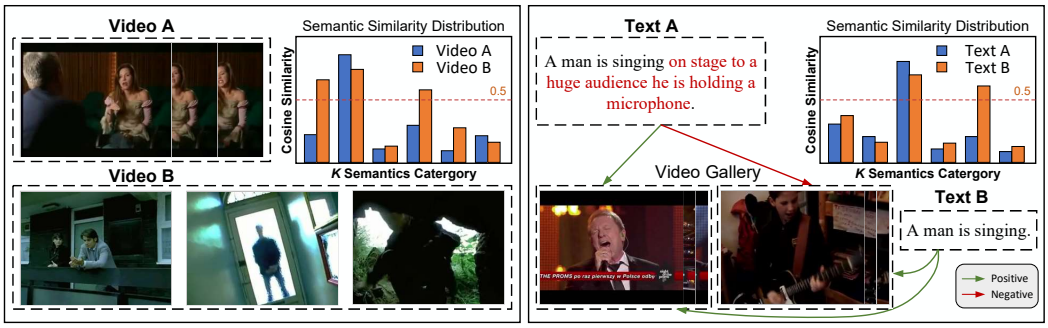
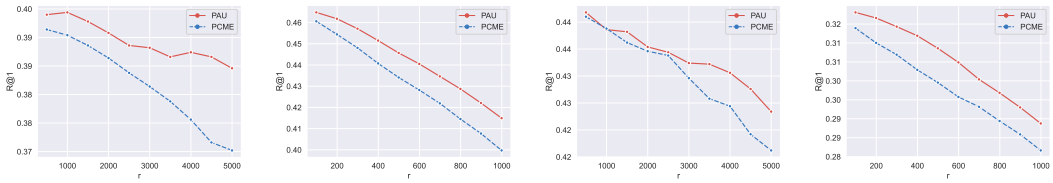


(a) Video modality (b) Text modality

Figure 7: **Illustration of confused matching in fickle videos and non-detailed texts.** Assuming the possible semantics of each modal subspace are finite with K categories. (a) A single-scene *Video A* can only match one semantics of "*talking*". By contrast, a multi-scene *Video B* can match to 3 semantics of "*talking*", "*shadow*", and "*cavern*". (b) *Text A* can only match the left video, while *Text B* with some details removed (in red) matches both videos.



(a) i2t on CLIP (b) t2i on CLIP (c) i2t on PCME (d) t2i on PCME

Figure 8: **The performance changes comparison after removing top-r instances with the highest uncertainty scores quantified by PCME and PAU on MS-COCO.** To fairly compare, We employ the removal on both predictions arising from CLIP and PCME. (a) and (b) show the performance changes on CLIP predictions. (c) and (d) show the performance changes on PCME predictions. In i2t, text instances are removed. In t2i, image instances are removed.

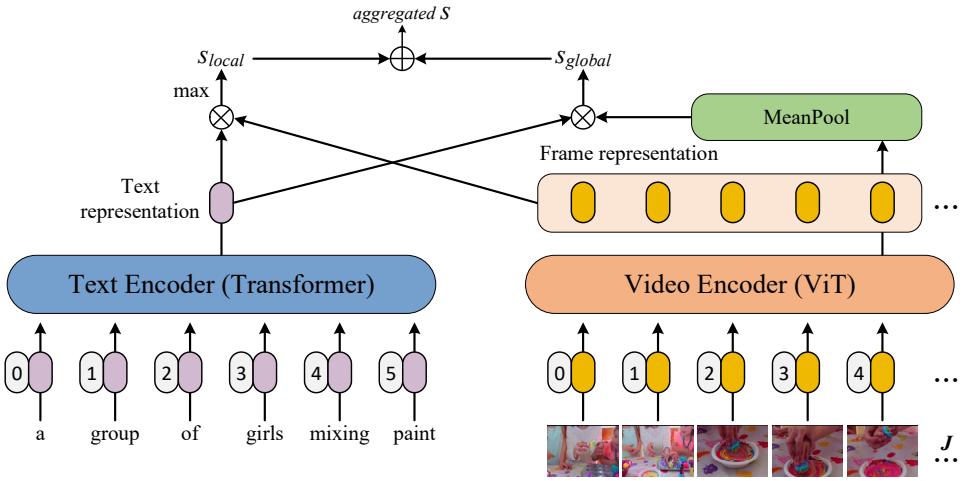


Figure 9: **The model structure of Baseline on video-text retrieval.** The model takes a video-text pair as input. For the input video, we encode it into J frame representations. For the input text, we encode it into a text representation. Then, a mean pooling layer aggregates J frame representations to a global representation, which calculates the global similarity s_{global} with the text representation. Meanwhile, the maximum value of the similarities between J frame representations and the text representation is selected as s_{local}. Finally, we take the average of s_{global} and s_{local} as the final similarity s, which will be constrained by a cross-entropy loss in the training process. $\otimes$ means cosine similarity. $\oplus$ means average addition.