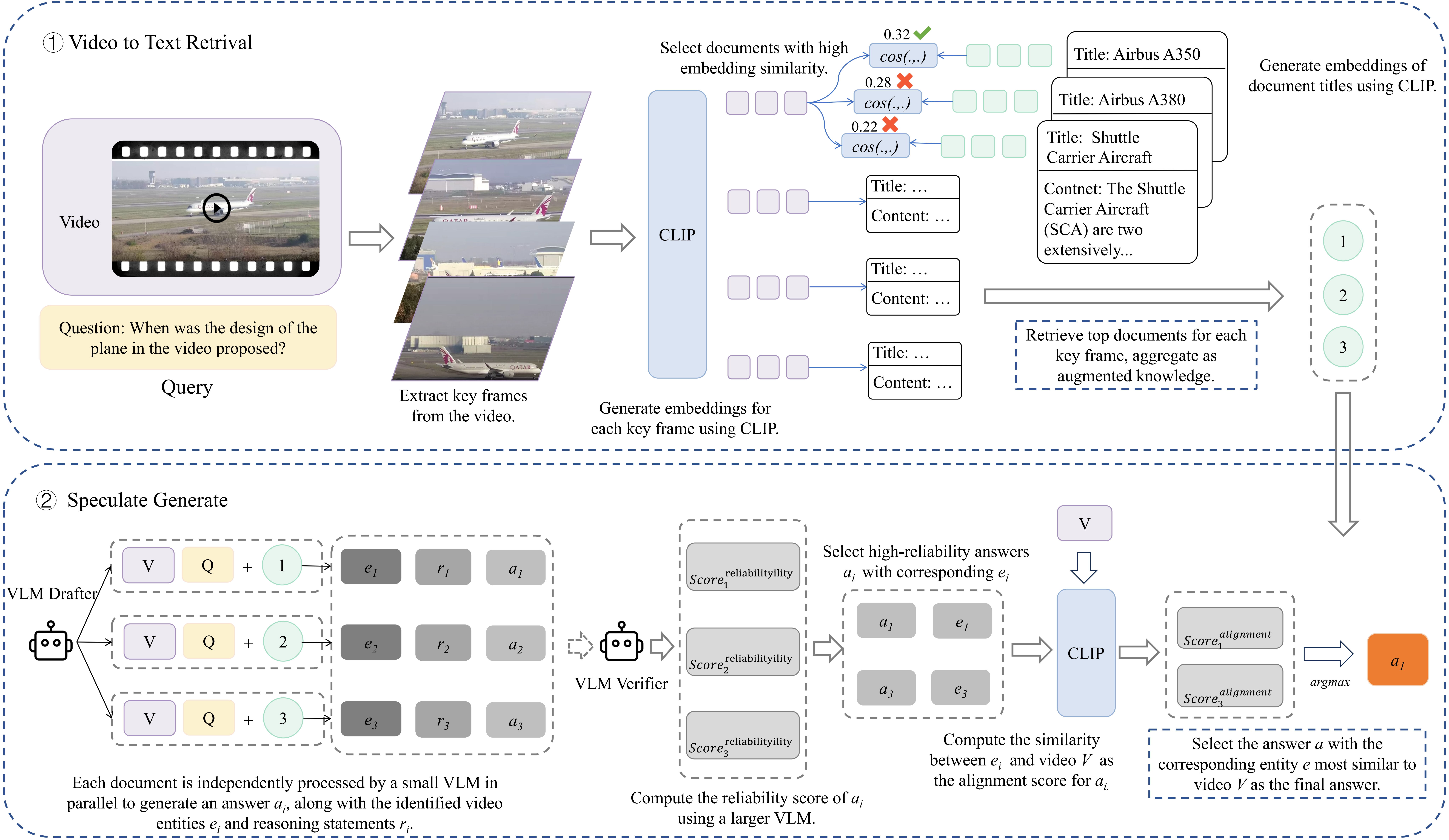

Select documents with high embedding similarity.

Generate embeddings of document titles using CLIP.

Extract key frames from the video.

Generate embeddings for each key frame using CLIP.

Retrieve top documents for each key frame, aggregate as augmented knowledge.

Each document is independently processed by a small VLM in parallel to generate an answer a_i, along with the identified video entities e_i and reasoning statements r_i.

Compute the reliability score of a_i using a larger VLM.

Select high-reliability answers a_i with corresponding e_i

Compute the similarity between e_i and video V as the alignment score for a_i.

Select the answer a with the corresponding entity e most similar to video V as the final answer.