GOAL: Text-to-video retrieval without access to manually-labeled videos, but allowing access to manually-labeled images.

INTRODUCTION

CAPTION SELECTION

• Caption selection: We caption 10 video frames with both ClipCap [1] and BLIP [2], then compute CLIPScore [4] (gray box) to select Top K=2 captions.

• Multi-caption query scoring (MCQS): We extend [5] to pool frame embeddings.

• Simple contrastive retrieval objective with the automatic video label.

TLDXR: We automatically label video frames with image captioners [1, 2] and select high-quality captions to train text-video retrieval models, pretrained on a text-image retrieval model, CLIP [3].

REFERENCES