VideoBERT: A Joint Model for Video and Language Representation Learning

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid

Google Research

Presented by

Md Mohaiminul Islam

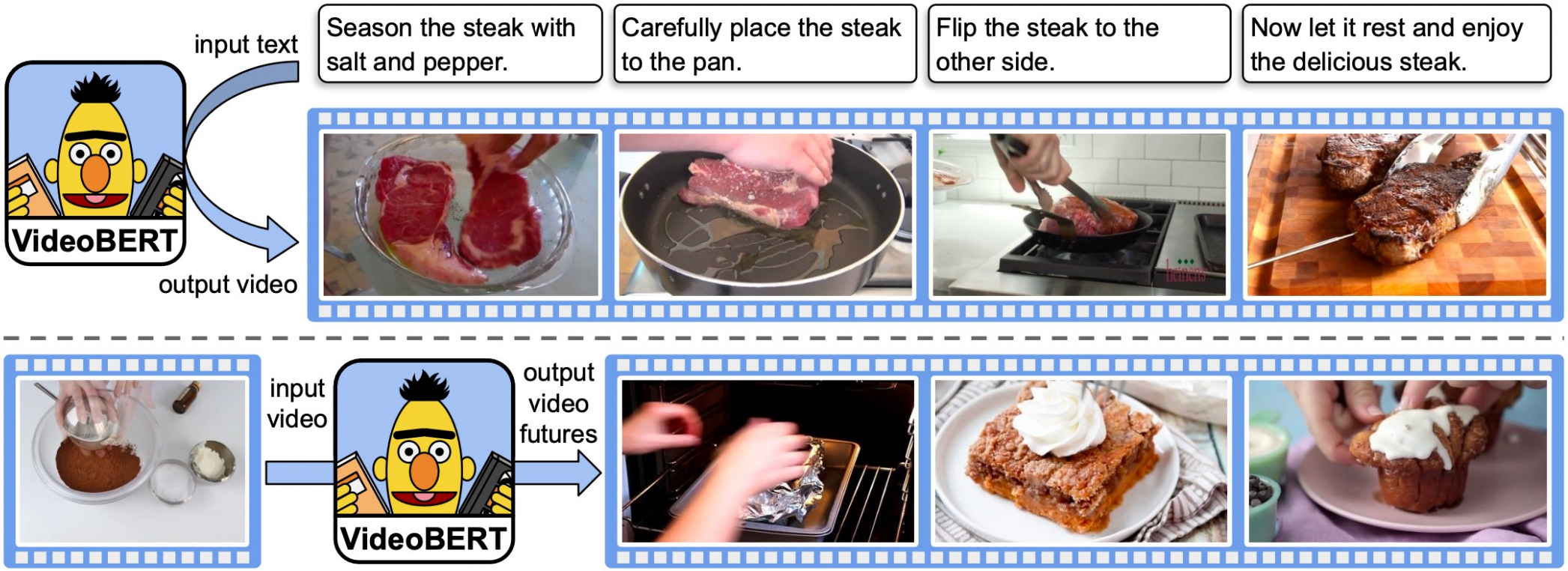
Motivation

- Self supervised to utilize unlabeled data
- Proxy task that can discover high-level semantic features

Approach

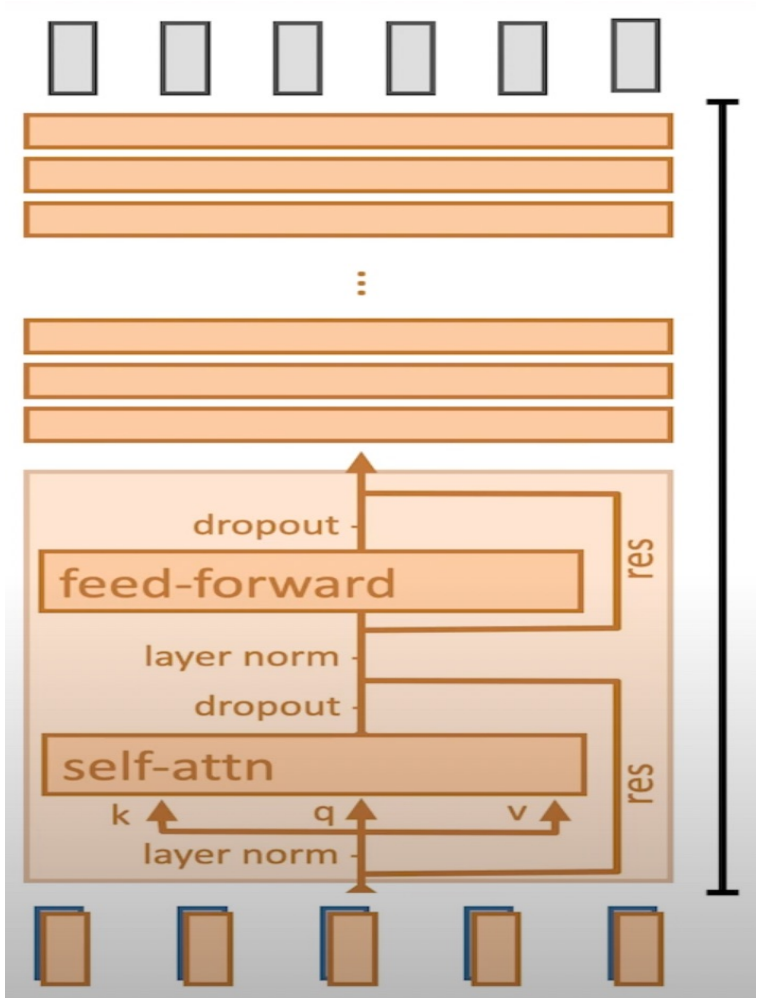
- Human language provides a natural source of “self” supervision.
- Model the relationship between the visual domain and the linguistic domain by:
  1. automatic speech recognition (ASR)
  2. vector quantization (VQ)
  3. BERT model
- apply BERT to learn a model of the form p(x, y)

Approach

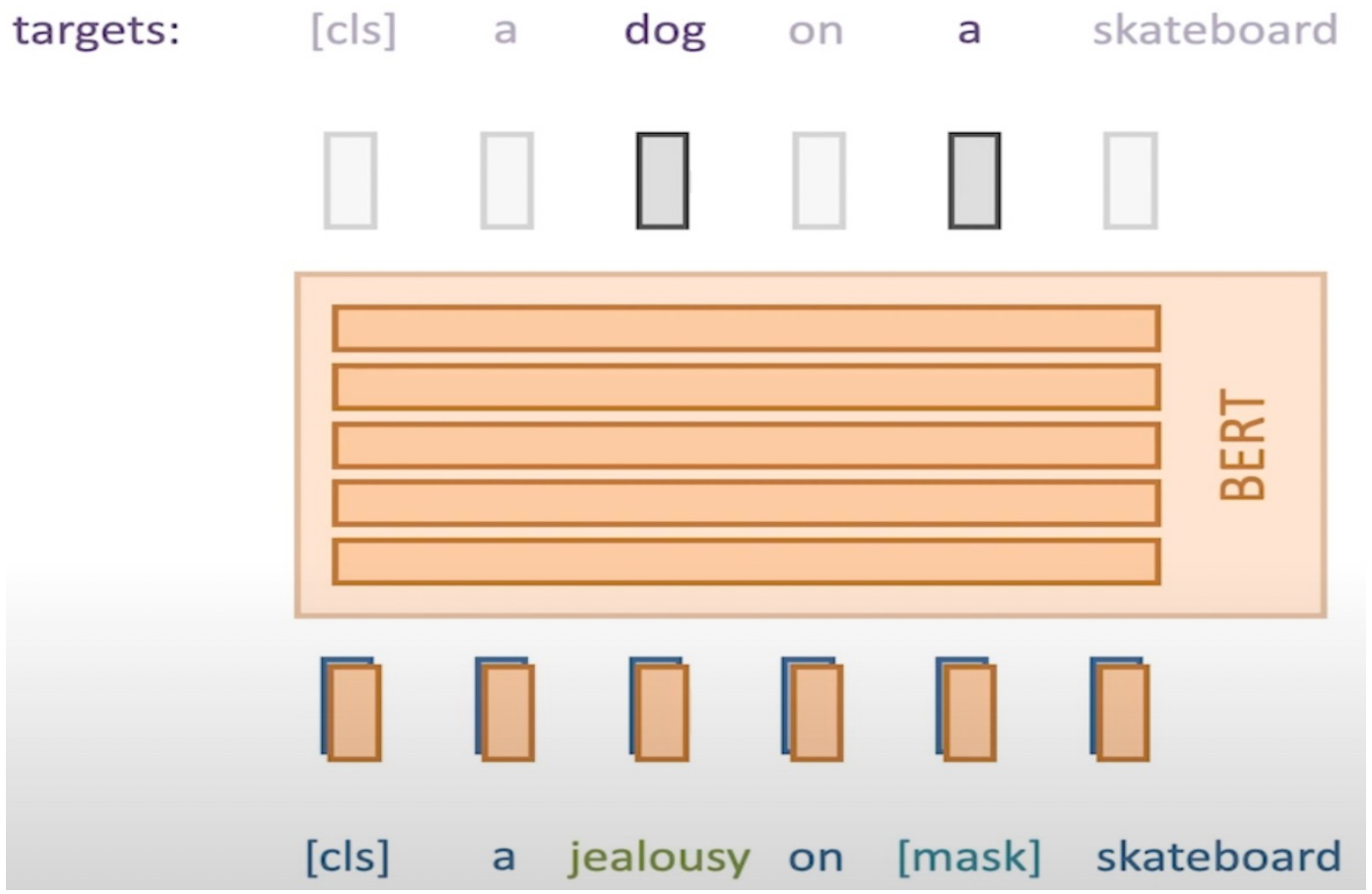

input text
Season the steak with salt and pepper.
Carefully place the steak to the pan.
Flip the steak to the other side.
Now let it rest and enjoy the delicious steak.
VideoBERT
output video
input video
VideoBERT
output video futures

The BERT model

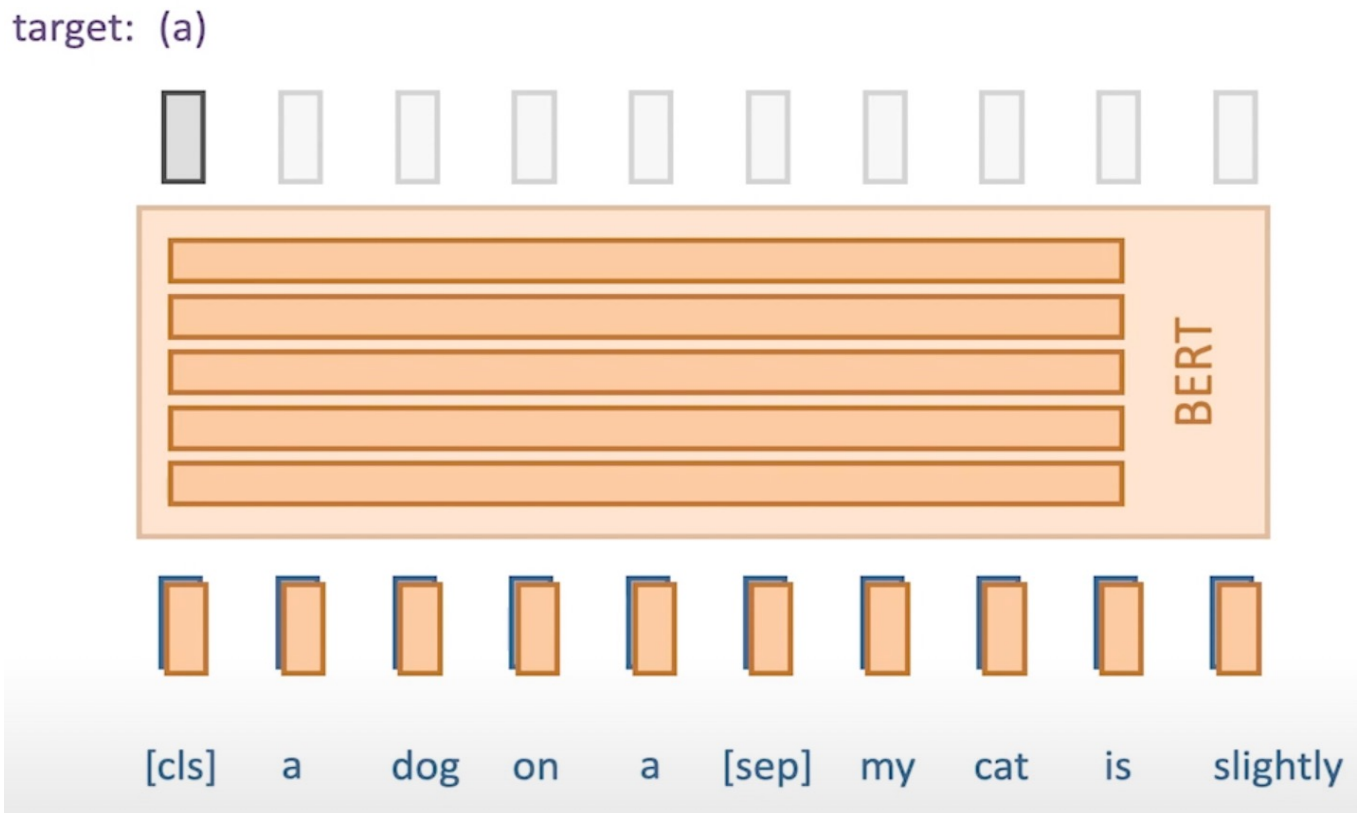
- Stack of transformer blocks



Task 1: Masking



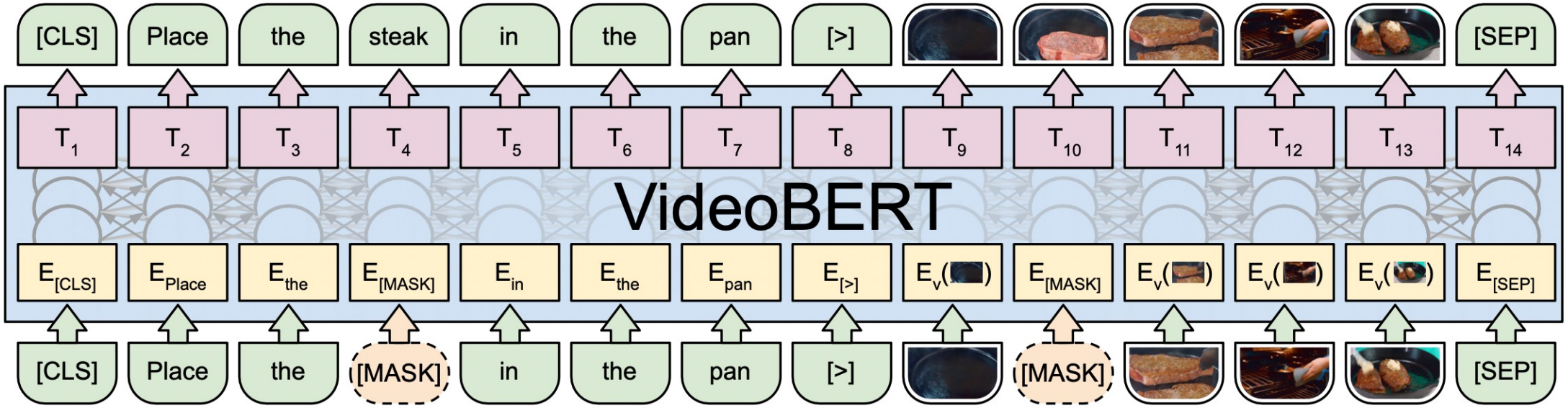
Task 2:



The VideoBERT model

BERT[CLS] let's make a traditional [MASK] cuisine [SEP] orange chicken with [MASK] sauce [SEP]

VideoBert: [CLS] orange chicken with [MASK] sauce [>] v01 [MASK] v08 v72 [SEP]



Linguistic-visual alignment task

Final hidden state of the [CLS] token to predict whether the linguistic sentence is temporally aligned with the visual sentence.

Three training regimes : text-only, video-only and video-text.

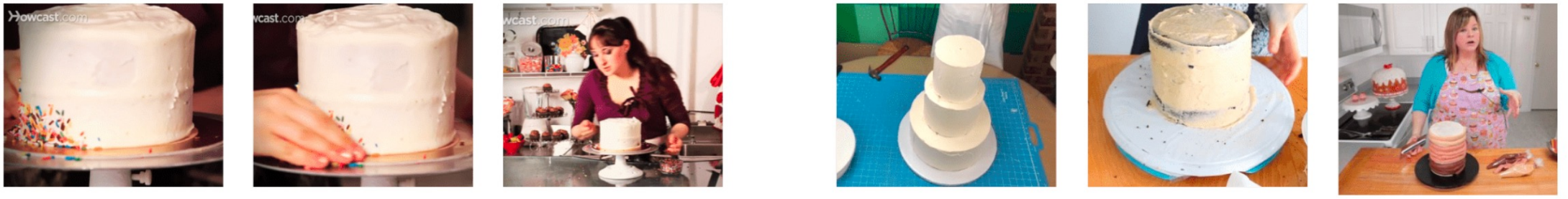
Dataset

- Instructional videos
- Publicly available cooking videos from YouTube
- 312K videos
- Text is obtained via YouTube’s automatic speech recognition (ASR)
- Evaluate on the YouCook II dataset

Video and Language Preprocessing

Video: S3D (pretrained on kinetics)

tokenize the visual features using hierarchical k- means



“but in the meantime, you're just kind of moving around your cake board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party.”

“apply a little bit of butter on one side and place a portion of the stuffing and spread evenly cover with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches.”

Zero-shot action classification YouCook II

now let me show you how to [MASK] the [MASK]

Top verbs: make, assemble, prepare
Top nouns: pizza, sauce, pasta

Method	Supervision	verb top-1 (%)	verb top-5 (%)	object top-1 (%)	object top-5 (%)
S3D [34]	yes	16.1	46.9	13.2	30.9
BERT (language prior)	no	0.0	0.0	0.0	0.0
VideoBERT (language prior)	no	0.4	6.9	7.7	15.3
VideoBERT (cross modal)	no	3.2	43.3	13.1	33.7

Benefits of large training sets

Method	Data size	verb top-1 (%)	verb top-5 (%)	object top-1 (%)	object top-5 (%)
VideoBERT	10K	0.4	15.5	2.9	17.8
VideoBERT	50K	1.1	15.7	8.7	27.3
VideoBERT	100K	2.9	24.5	11.2	30.6
VideoBERT	300K	3.2	43.3	13.1	33.7

Transfer learning for captioning

now let's [MASK] the [MASK] to the [MASK], and then [MASK] the [MASK]

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou *et al.* [39]	7.53	3.84	11.55	27.44	0.38
S3D [34]	6.12	3.24	9.52	26.09	0.31
VideoBERT (video only)	6.33	3.81	10.81	27.14	0.47
VideoBERT	6.80	4.04	11.01	27.50	0.49
VideoBERT + S3D	**7.59**	**4.33**	**11.94**	**28.80**	**0.55**

Qualitative results


GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

S3D: cut the tomatoes into thin slices


GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan

Review

Strengths:

- A simple and effective approach.
- Vector quantization method enables the model to learn high-level semantic features.
- A large-scale multimodal dataset of instructional videos.

Weaknesses:

- Limited experiments.

Discussion

- How would you compare the model with VAE or GAN? What are the advantages or disadvantages?
- Do you think it is a self-supervised approach?