Video-and-Language Pre-training

Luowei Zhou

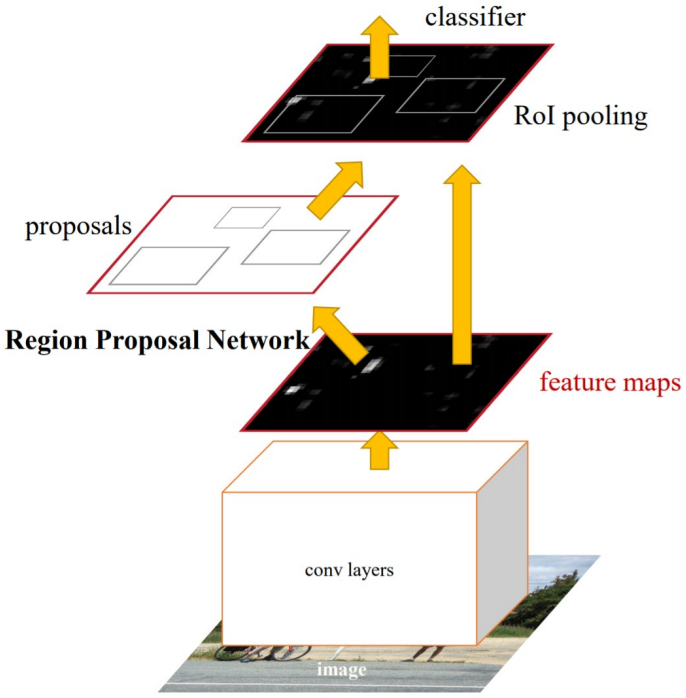
06/20/2021

Outline

- Data as fuel – The rise of pre-training data
- Method Overview and Taxonomy
- Reconstructive Methods and Contrastive Methods
- Video-Language-*Audio* – The new favorite?
- From image to video and back
- Downstream Tasks and Results
- Video-And-Language Understanding Evaluation (VALUE) benchmark
- Conclusion

Pre-training isn't new

- In fact, it is rather pervasive!



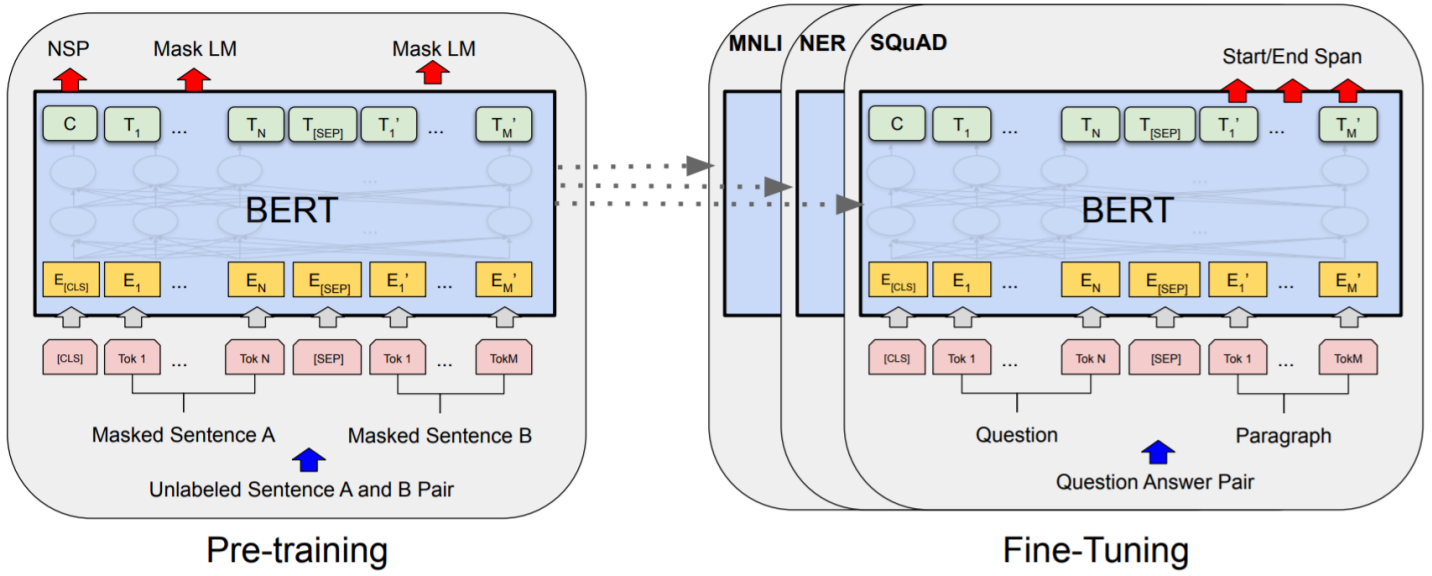


Figure credits:
Ren et al., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI 2016.
Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

Pre-training isn't new

- This has inspired a series of work at the intersection of image and language, thanks to the availability of large high-quality curated datasets (e.g., COCO, Conceptual Captions).

	In-domain		Out-of-domain	
Split	COCO Captions	VG Dense Captions	Conceptual Captions	SBU Captions
train	533K (106K)	5.06M (101K)	3.0M (3.0M)	990K (990K)
val	25K (5K)	106K (2.1K)	14K (14K)	10K (10K)

Table 1: Statistics on the datasets used for pre-training. Each cell shows #image-text pairs (#images)

- But not so much in the video domain.

Table credit: Chen et al., UNITER: UNiversal Image-TExt Representation Learning. ECCV 2020.

Pre-training isn’t new

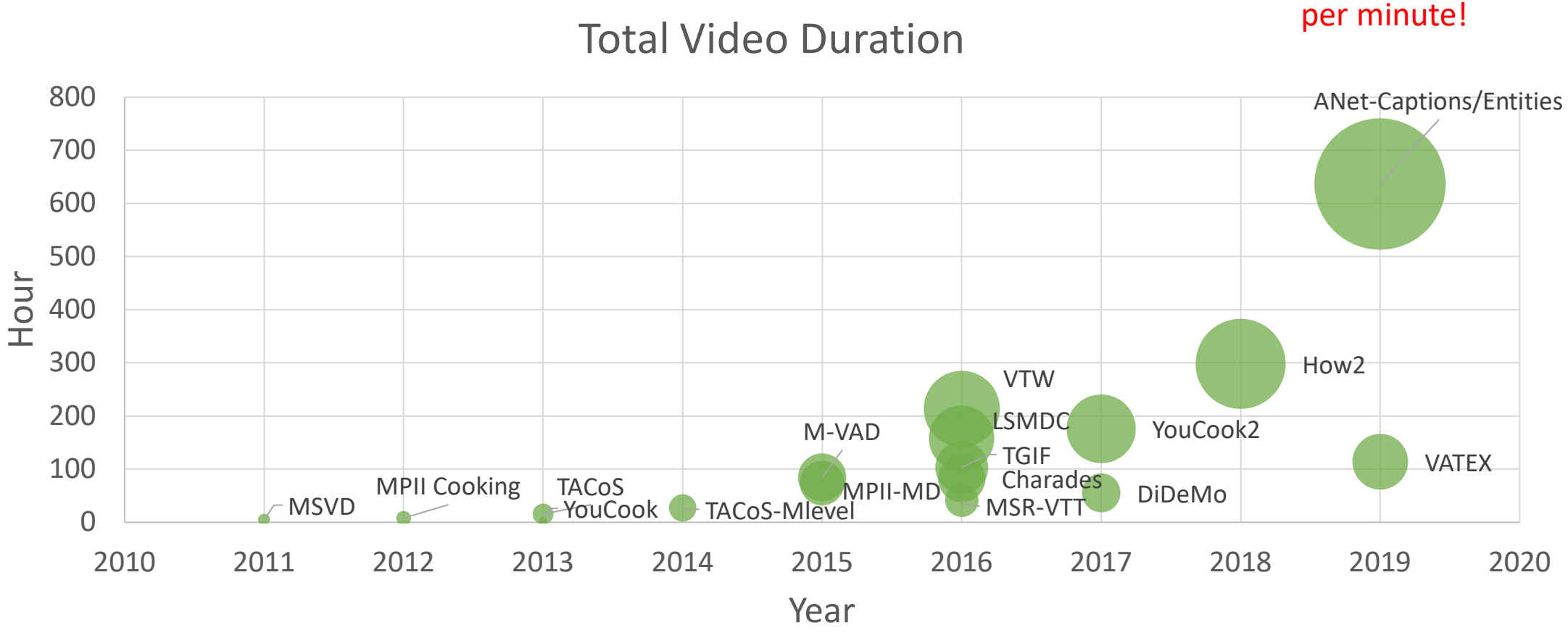
- The reasons why the video-and-language field has been lagging behind is mainly due to:

- The challenge in harvesting large-scale data;

- The challenge in annotating those data.

Evolution of video-language datasets

As a comparison, 500hr worth of videos are uploaded to YouTube per minute!



Video credit: COIN dataset

The Era of Pre-training

- “Free” annotations become accessible (i.e., subtitles or ASR transcripts)

Transcript

11:27	it's cooled down now so just mix it up a
11:29	bit then we brush it on and the star of
11:38	the show next we're gonna put the rest
11:40	of the scallions on here put it nice and
11:42	**evenly all around we're going to roll it**
11:48	onto itself and so we start at one side
11:51	and we roll it onto itself here okay
11:55	until you have a log try and keep it
12:00	tight
12:01	[Music]
12:10	and so I'm gonna cut this into four half

English (auto-generated)

Figure credit: Making Scallion Pancake Beef Rolls: https://www.youtube.com/watch?v=vTmgLKtx49Y

Video-and-Language Pre-training

- Paired video clips and subtitles

"Keep rolling tight and squeeze the air out to its side and you can kind of pull a little bit."

Keep rolling tight and squeeze the air out to its side and you can kind of pull a little bit.

- The resulted datasets are magnitudes bigger!

Figure credit: https://ai.googleblog.com/2019/09/learning-cross-modal-temporal.html

Pre-training Data

- The major video-and-language dataset for pre-training:

HowTo100M Dataset

[Miech et al., ICCV 2019]

- 1.22M instructional videos from YouTube
- Each video is 6 minutes long on average
- Over 100 million pairs of video clips and associated narrations

Pre-training Data

- Emerging public video-and-language datasets for pre-training:

TV Dataset

[Lei et al., EMNLP 2018]

- 22K video clips from 6 popular TV shows
- Each video clip is 60-90 seconds long
- Dialogue (“character: subtitle”) is provided

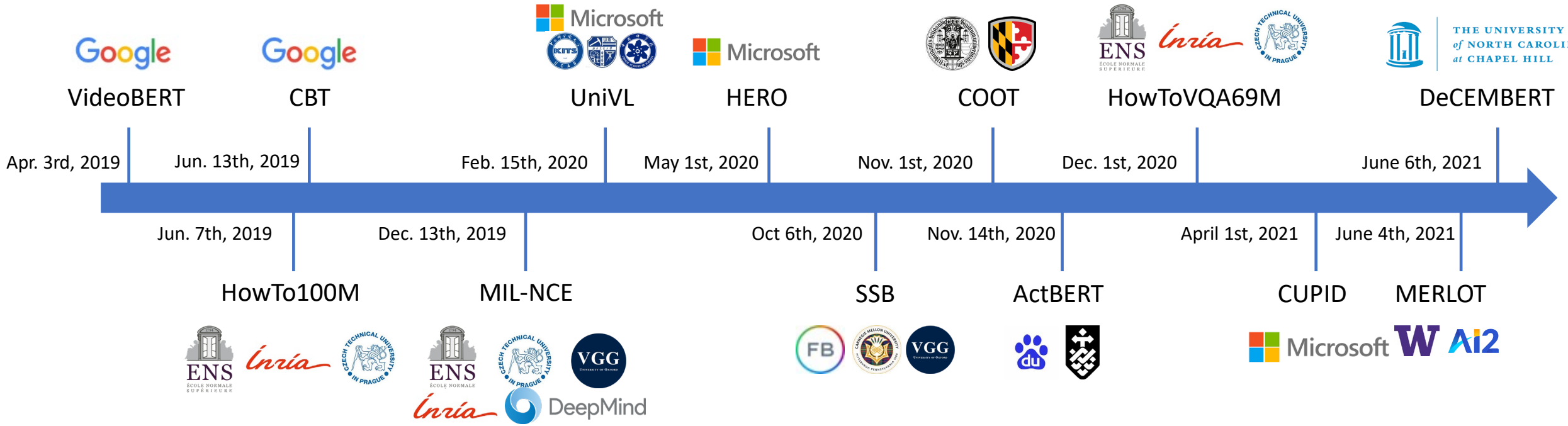
Auto-captions on GIF Dataset

[Pan et al., arXiv 2020]

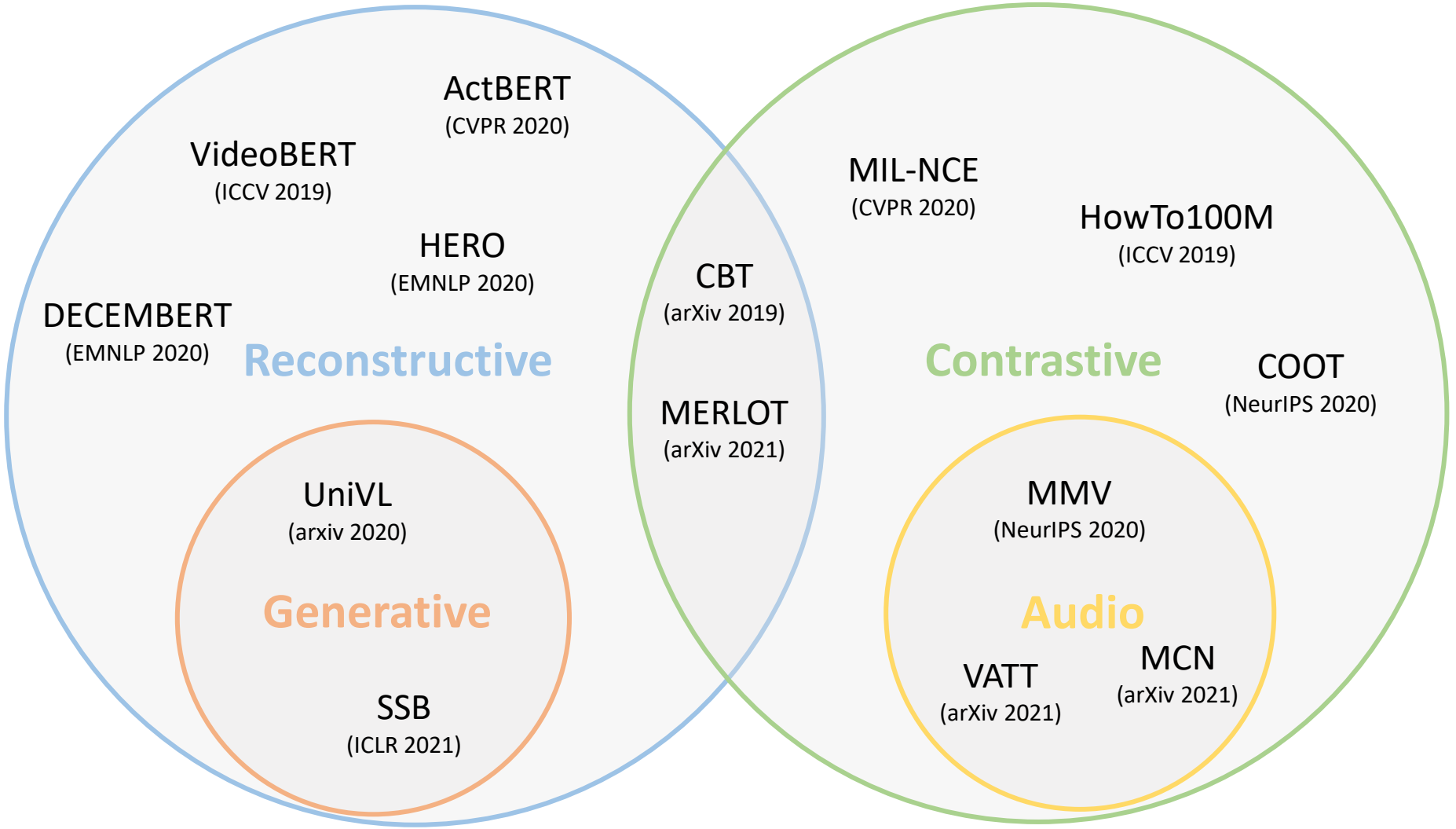
- 163K GIFs automatically crawled from web
- Each GIF is a few seconds long
- Cover a variety of categories

Figure credits: from the original papers

Method Overview

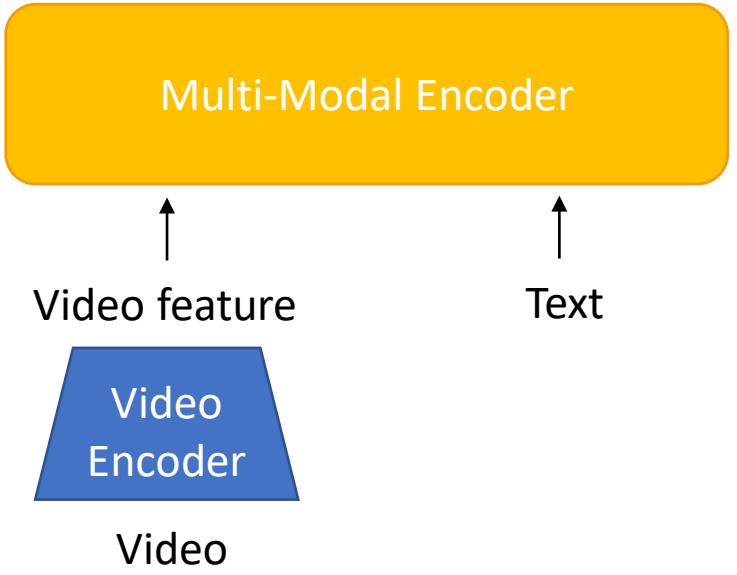


Taxonomy


ActBERT
(CVPR 2020)
VideoBERT
(ICCV 2019)
HERO
(EMNLP 2020)
DECEMBERT
(EMNLP 2020)
Reconstructive
UniVL
(arxiv 2020)
Generative
SSB
(ICLR 2021)
CBT
(arXiv 2019)
MERLOT
(arXiv 2021)
MIL-NCE
(CVPR 2020)
HowTo100M
(ICCV 2019)
Contrastive
COOT
(NeurIPS 2020)
MMV
(NeurIPS 2020)
Audio
VATT
(arXiv 2021)
MCN
(arXiv 2021)

Reconstructive Methods

- BERT-inspired; usually adopt the early fusion architecture.



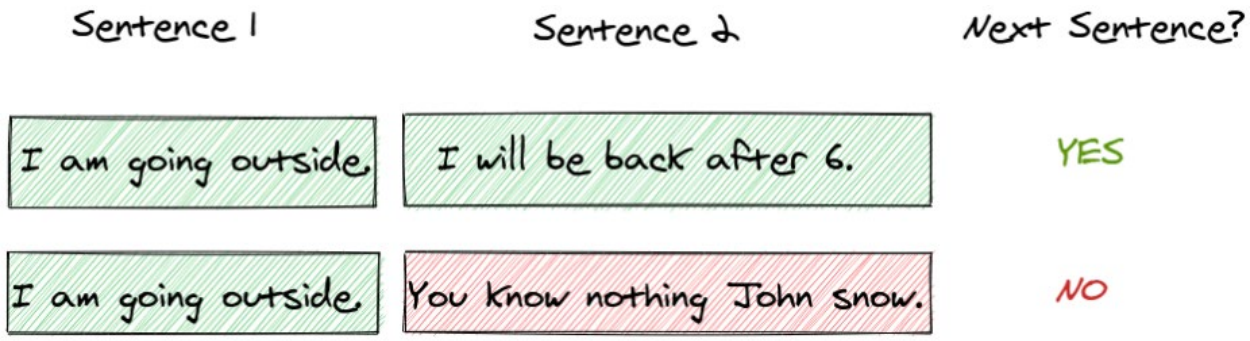
- Usually leverage pre-trained unimodal feature/backbone (e.g., BERT, I3D)
- Image counterparts: ViLBERT/VLP/UNITER/OSCAR

Figure credit: Sun et al., VideoBERT: A Joint Model for Video and Language Representation Learning. ICCV 2019.

Background (BERT)

- BERT – Bidirectional Encoder Representations from Transformers

- Training Objectives
  - Masked Language Modeling (MLM)
  - Next Sentence Prediction (NSP)



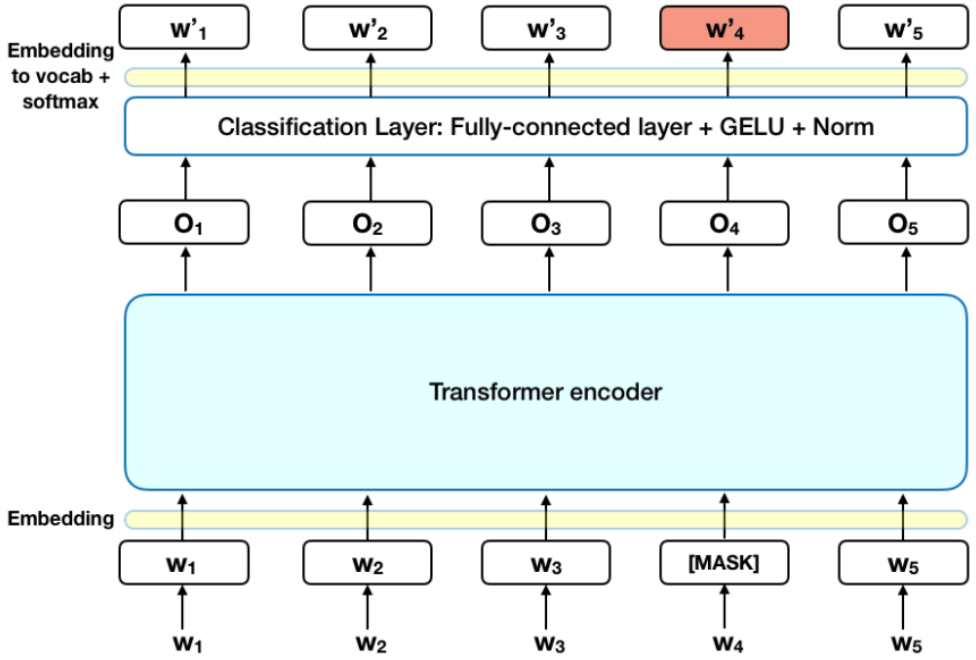
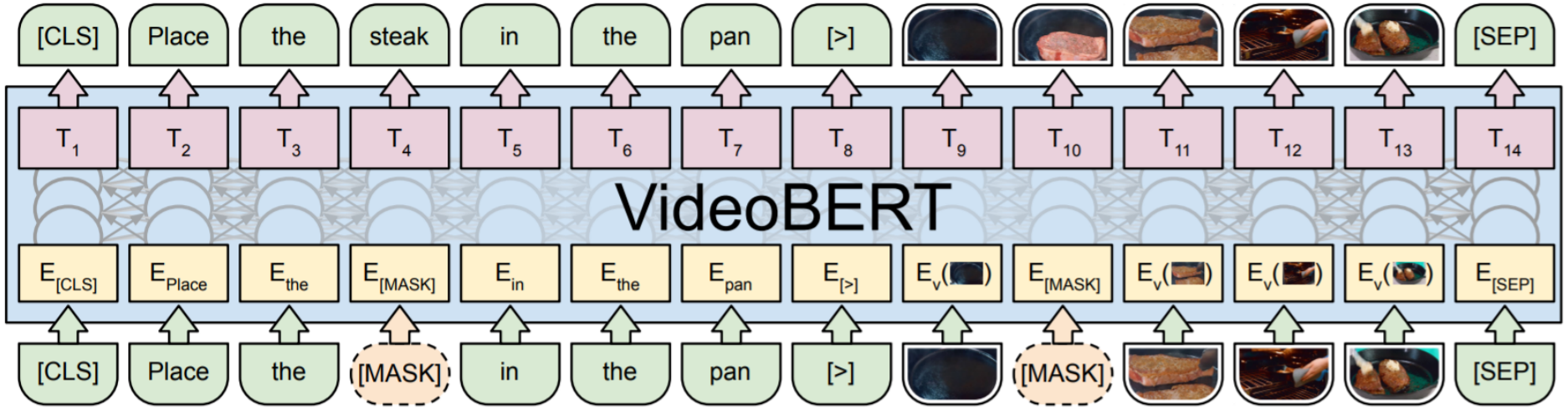


Figure credits: https://www.kdnuggets.com/2018/12/bert-sota-nlp-model-explained.html
https://amitness.com/2020/02/albert-visual-summary/

VideoBRET

- Pre-training: 312K cooking videos from YouTube
- Video feature: Kinetics-pretrained S3D; then tokenize into 21K clusters using hierarchical K-means. Multi-Modal Encoder: BERT-large.
- Objectives: Masked Language Modeling (MLM), Masked Frame Modeling (MFM), Video-Text Matching (VTM)



Sun et al., VideoBERT: A Joint Model for Video and Language Representation Learning. ICCV 2019.

VideoBRET

- Adding more data generally gives better results

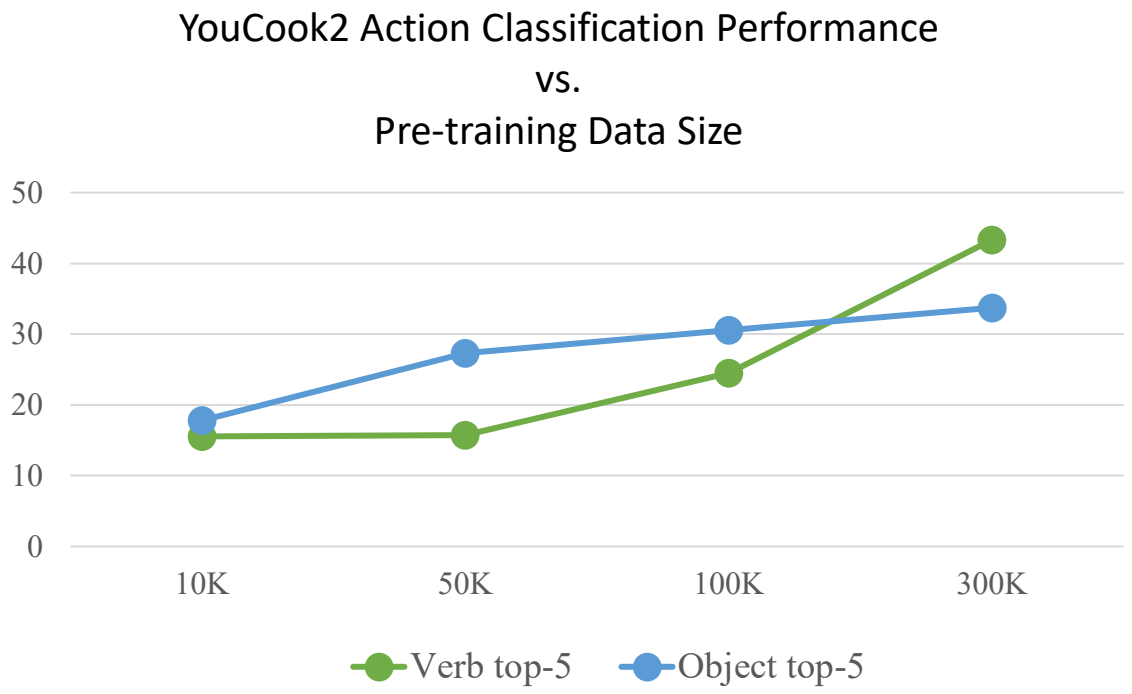
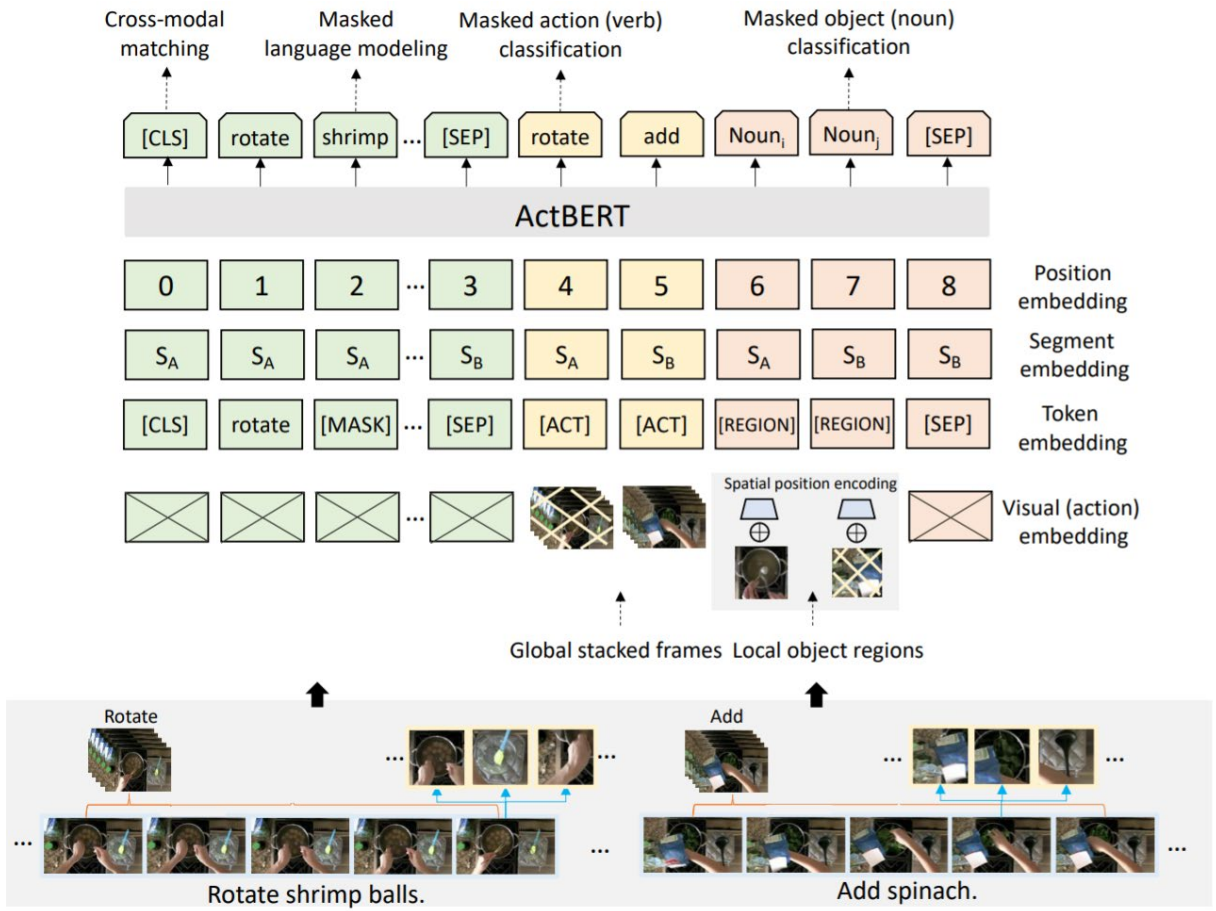


Figure credit: https://rohit497.github.io/Recent-Advances-in-Vision-and-Language-Research/slides/tutorial-part5-pretraining.pdf

ActBERT

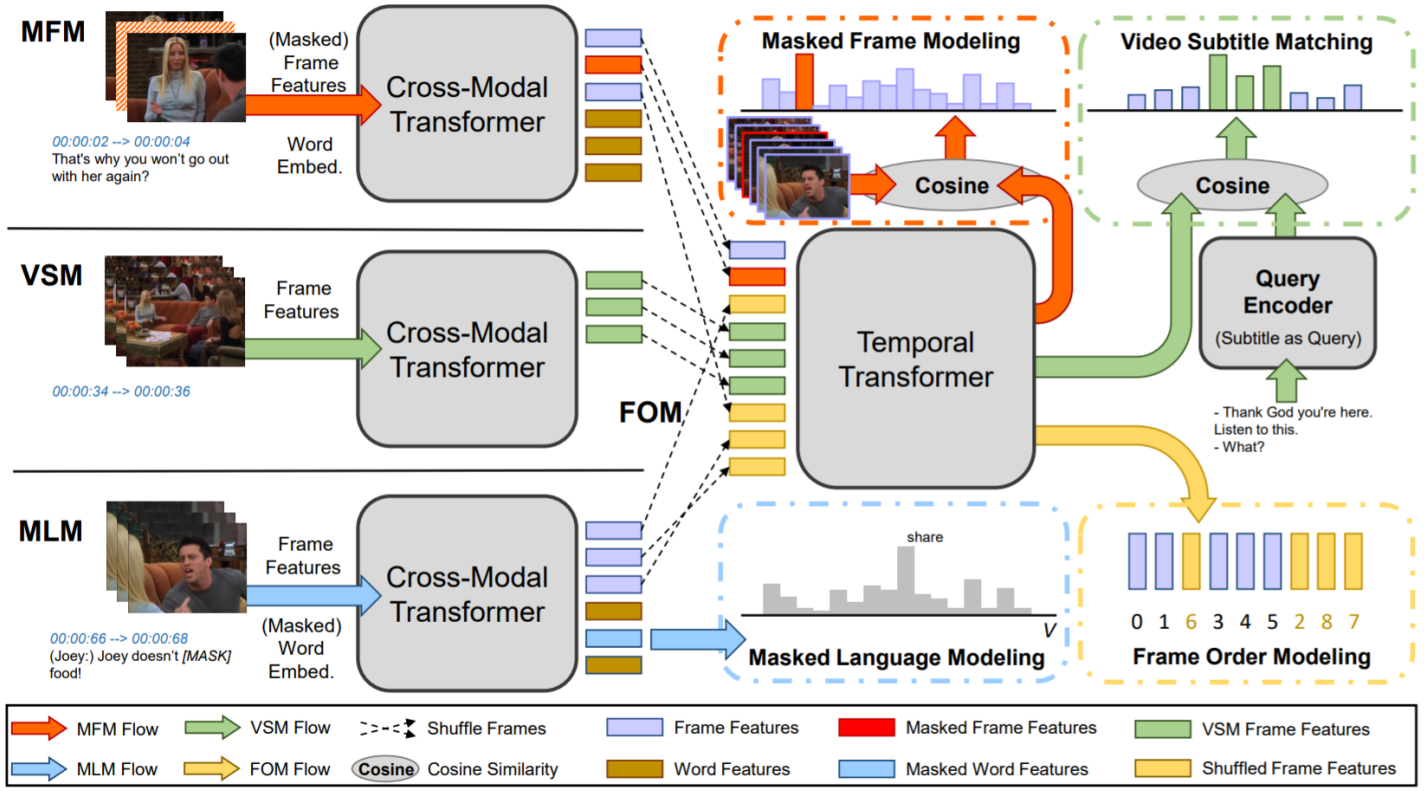
- Pre-training: HowTo100M
- Video feature: object region feature from Faster RCNN; Kinetics-pretrained R(2+1)D.
- Multi-Modal Encoder: BERT-base.
- Training objectives
  - MLM, VTM
  - Masked Object (Noun) Classification
  - Masked Action (Verb) Classification



Zhu et al., ActBERT: Learning Global-Local Video-Text Representations. CVPR 2020.

HERO (**H**ierarchical **E**ncode**r** for **O**mni-representation learning)

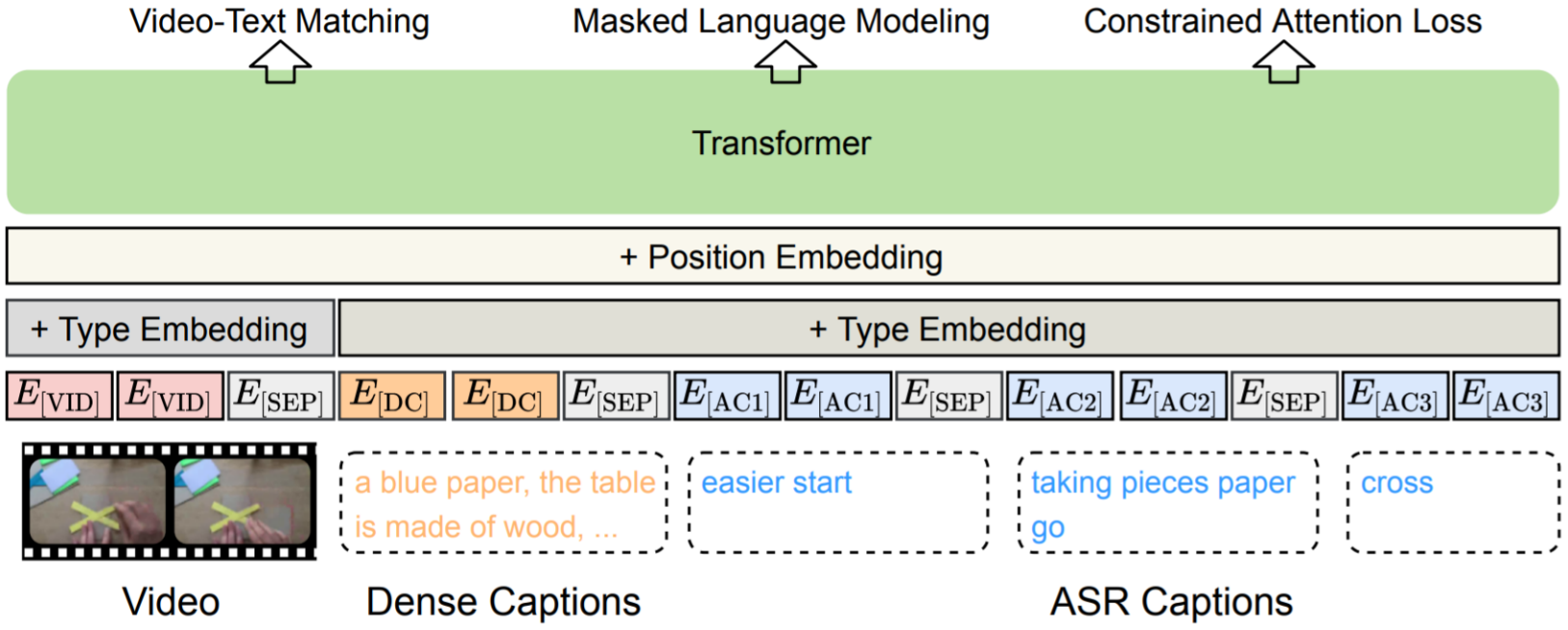
- Objectives: MLM, MFM; New: Video-Subtitle Matching (VSM), Frame Order Modeling (FOM)



Li et al., HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. ECCV 2020.

DECEMBERT (**De**nse **C**aptions and **E**ntropy **M**inimization)

- Dense captions input (from a VG pre-trained dense captioning model)
- Attention Entropy Minimization (deal with the misalignment issue between video clip and subtitle through sharp attention).



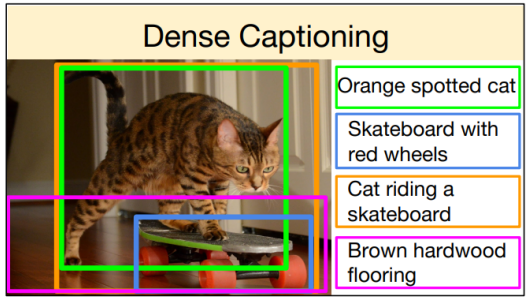
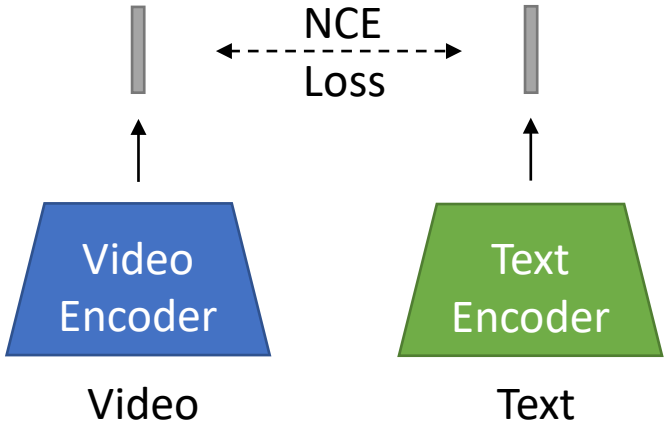


Figure credit: Johnson et al., CVPR 2016.

Tang et al., DECEMBERT: Learning from Noisy Instructional Videos via Dense Captions and Entropy Minimization. NAACL 2021.

Contrastive Methods

- Contrastive learning-inspired
- Usually adopt the late fusion architecture:



- Usually trained from scratch to learn a general feature representation
- Image counterpart: CLIP

GIF credit: https://docs.google.com/presentation/d/1ccddJFD_j3p3h0TCqSV9ajSi2y1yOfh0-lJoK29ircs/edit#slide=id.g8c1b8d6efd_0_17

Background (Contrastive Learning)

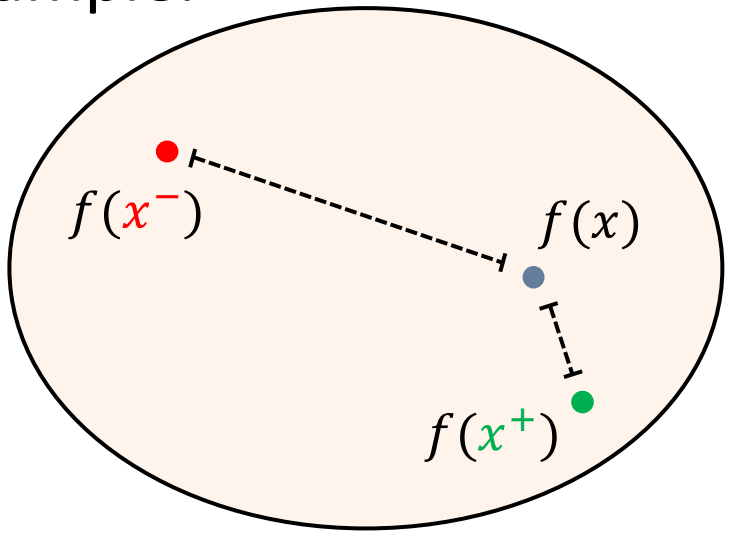
- Given a data point x, contrastive methods aim to learn an encoder f such that:

$$\mathrm{S}\big(f(x), f(x^+)\big) \gg S\big(f(x), f(x^-)\big),$$

- where x^+ is a data point similar to x, referred to as a *positive* sample, x^- is dissimilar to x, referred to as a *negative* sample.
- The score function S could simply be vector inner product (or cosine similarity).

$$S\big(f(x), f(x^+)\big) = f(x)^{\mathrm{T}} f(x^+)$$

- Most of the work until now is on how to define *positive* & *negative* samples.

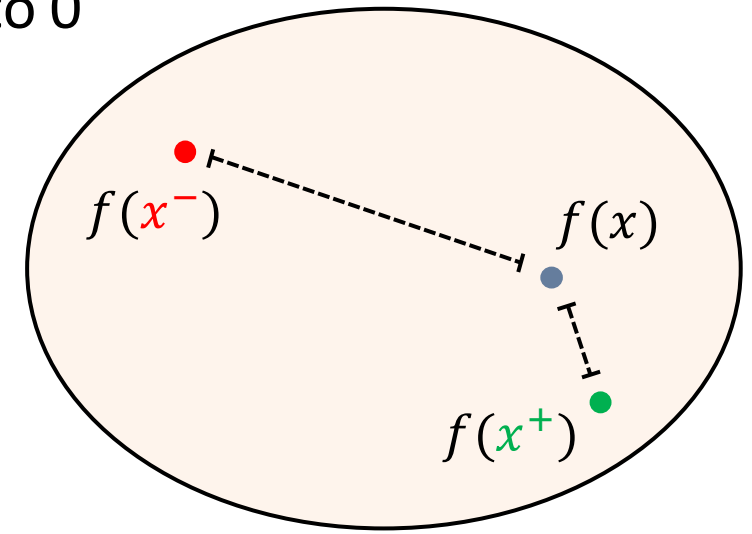


Most of content from this section is borrowed from https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html

Background (Contrastive Learning)

- Based on the objective function, contrastive methods fall into three categories.

- Logistic Loss (e.g., the VTM/NSP objective)
  - Regress $S\big(f(x), f(x^+)\big)$ to 1 and $S\big(f(x), f(x^-)\big)$ to 0



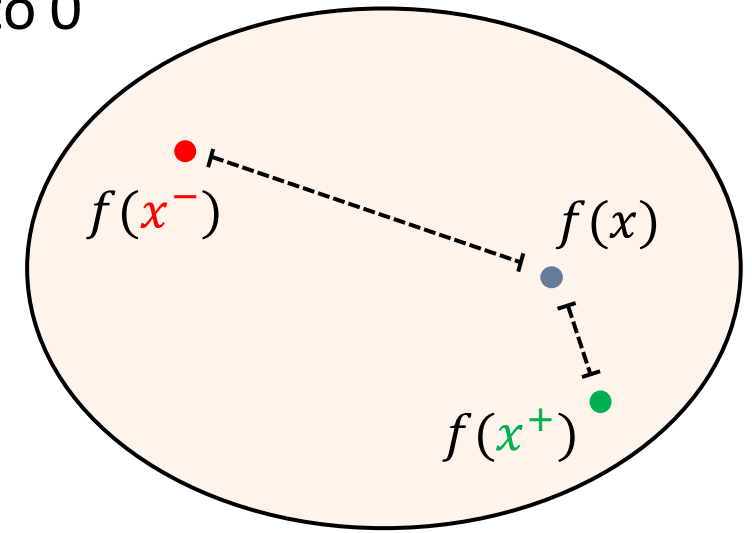
Background (Contrastive Learning)

- Based on the objective function, contrastive methods fall into three categories.

- Logistic Loss (e.g., the VTM/NSP objective)
  - Regress $S\big(f(x), f(x^+)\big)$ to 1 and $S\big(f(x), f(x^-)\big)$ to 0

- Margin Loss (e.g., see later in COOT)
  - Minimize the total hinge loss:

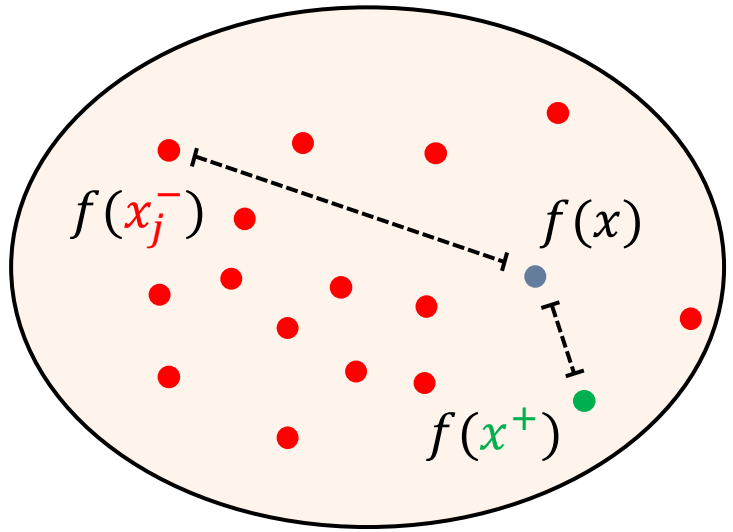
$$\max(S\big(f(x), f(x^-)\big) - S\big(f(x), f(x^+)\big) + \Delta, 0)$$



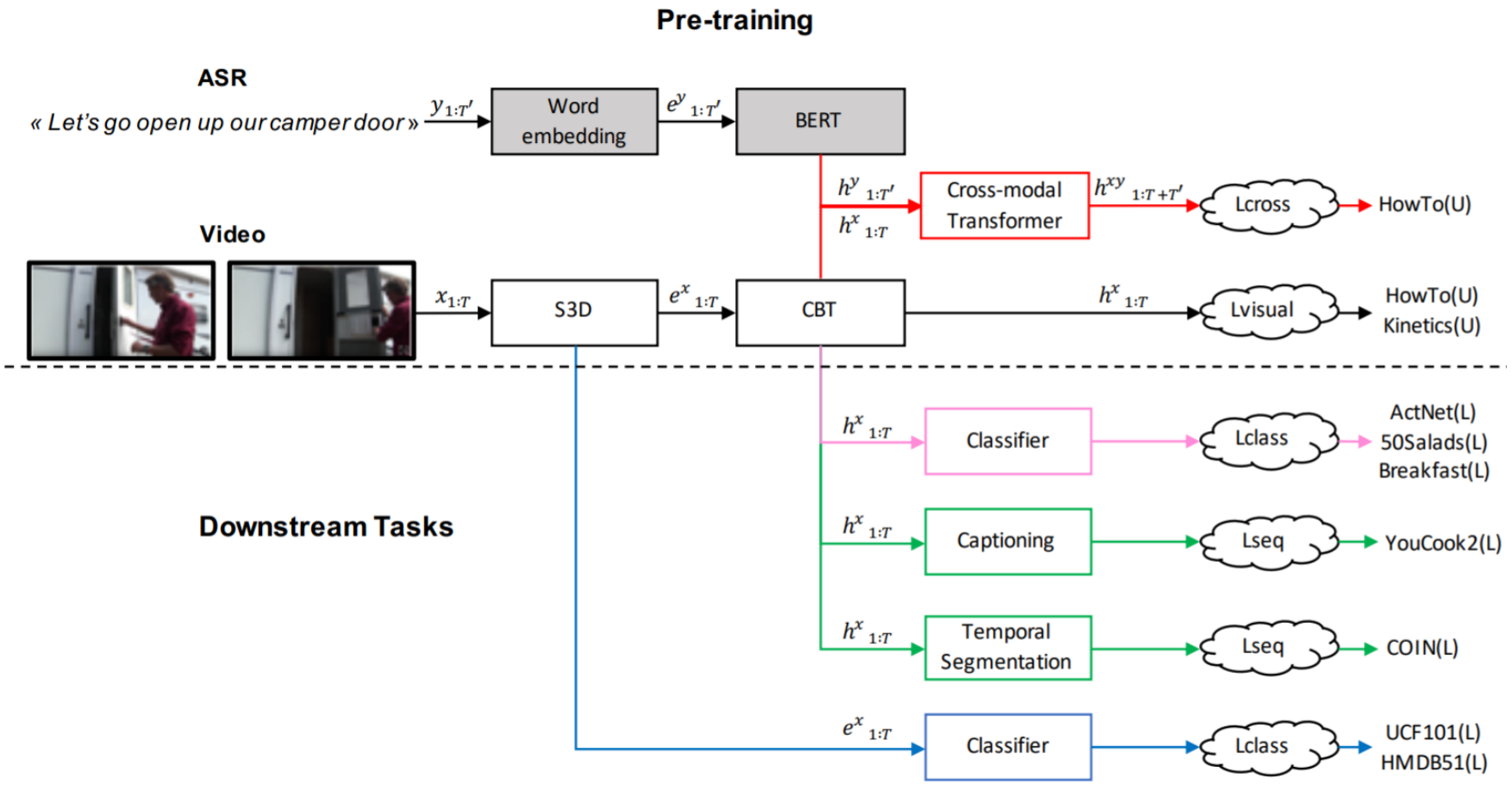
Background (Contrastive Learning)

- Noise-Contrastive Estimation (NCE) Loss
  - Use all other samples from the minibatch as negative samples
  - Cross entropy loss on an N-way Softmax classifier

$$-\log \frac{\exp(S(f(x), f(x^+))}{\exp(S(f(x), f(x^+)) + \sum_j \exp(S(f(x), f(x_j^-))}$$



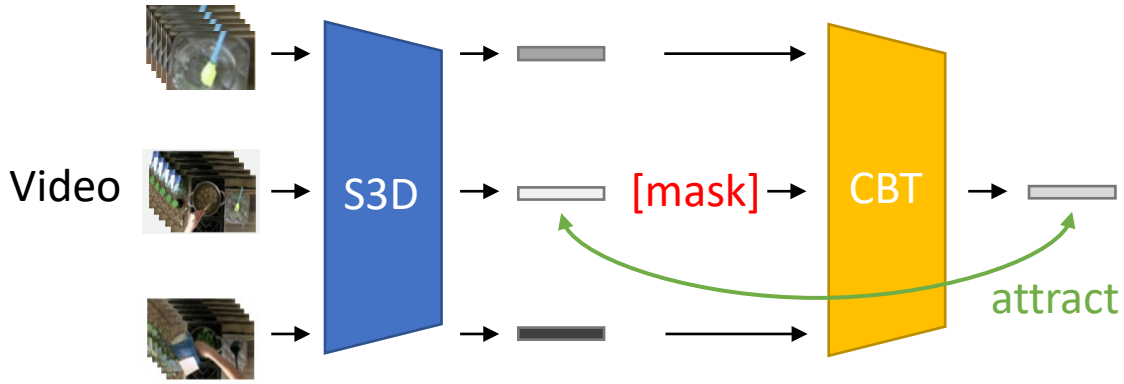
CBT: **C**ontrastive **B**idirectional **T**ransformer



Sun et al., Learning Video Representations using Contrastive Bidirectional Transformer. arXiv 2019.

CBT: Contrastive Bidirectional Transformer

- Objectives: i) Video NCE and ii) Video-Language NCE (VL-NCE).
- VL-NCE is simple, any paired clip and subtitle are considered a positive pair and the rest of the clips/subtitles in the minibatch are negatives.
- For Video NCE:



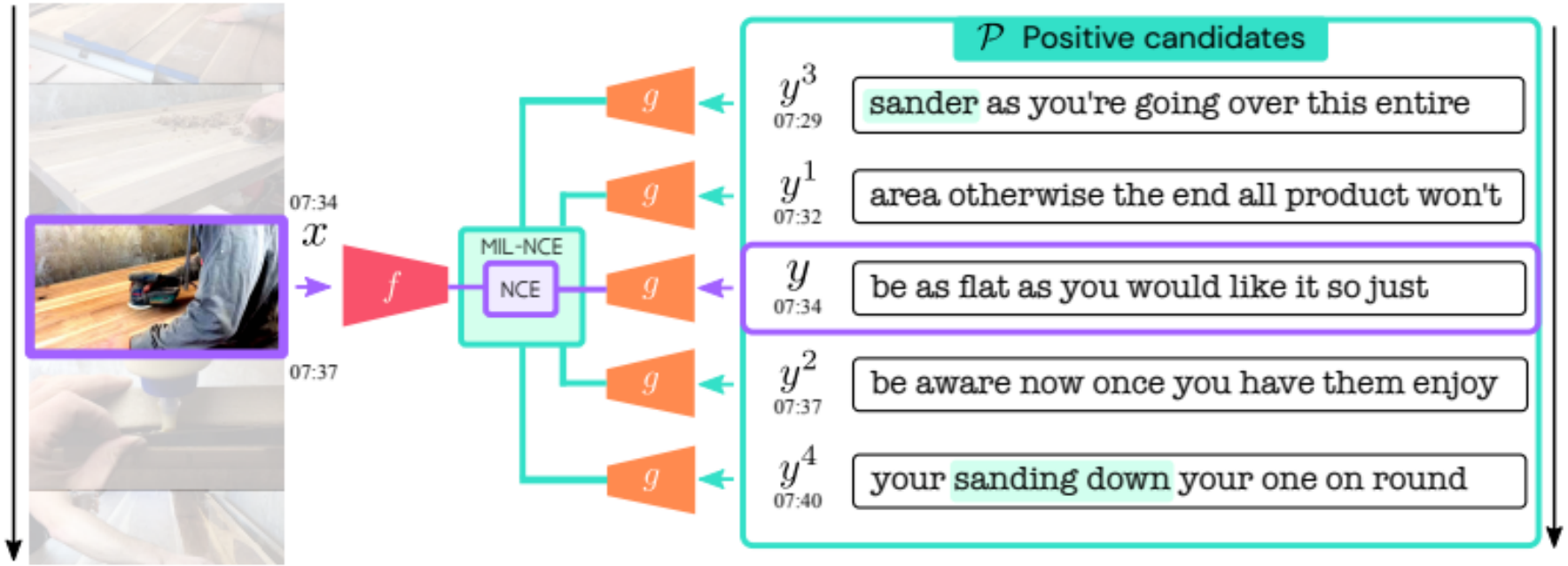
CBT is a shallow
(2-layer) Transformer

- A similar objective is used in HERO (MFM with NCE).

Sun et al., Learning Video Representations using Contrastive Bidirectional Transformer. arXiv 2019.

MIL-NCE

- It uses VL-NCE, with a twist on multiple instance learning (MIL) to address the misalignment issue between video clip and subtitle.

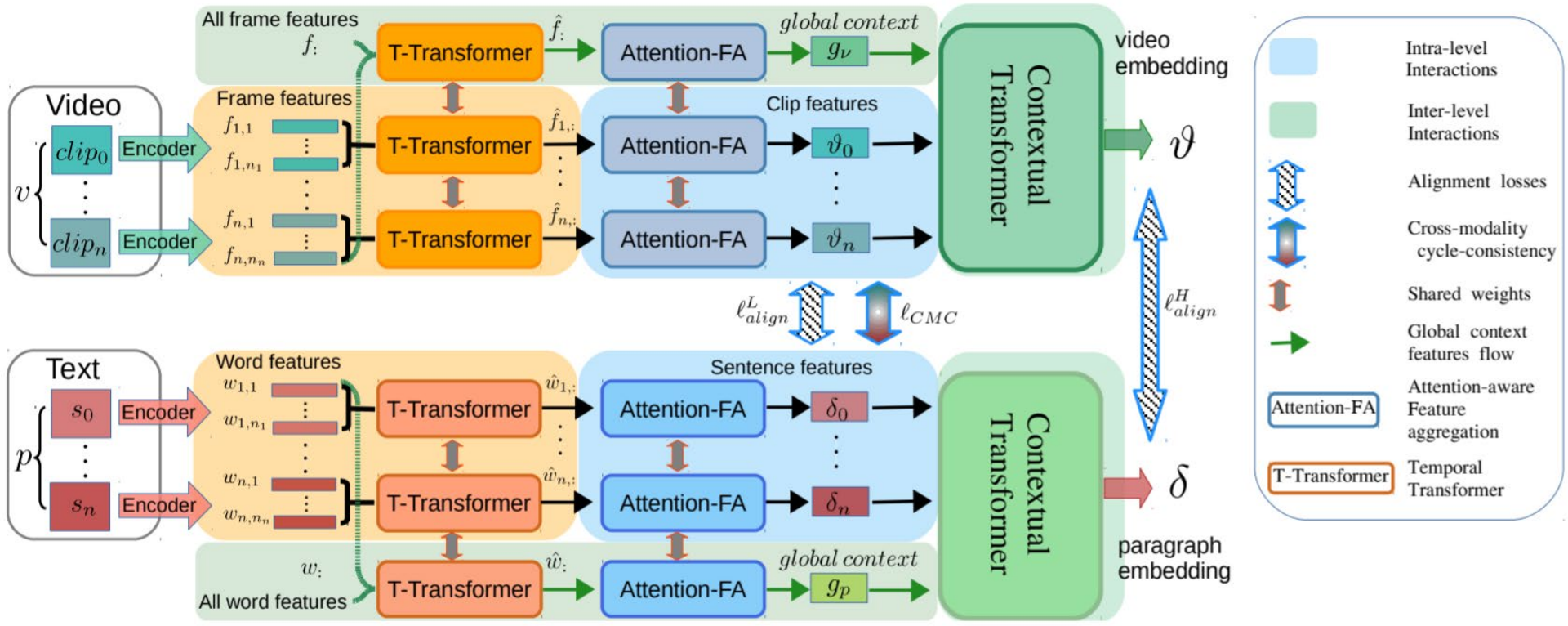


(a) **Examples of positive candidates**

Miech et al., End-to-End Learning of Visual Representations from Uncurated Instructional Videos. CVPR 2020.

COOT (**Coo**perative hierarchical **T**ransformer)

- Margin loss on clip-level and video-level alignment



Ging et al., COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. NeurIPS 2020.

COOT (**Coo**perative hierarchical **T**ransformer)

- Cross-modality cycle-consistency loss

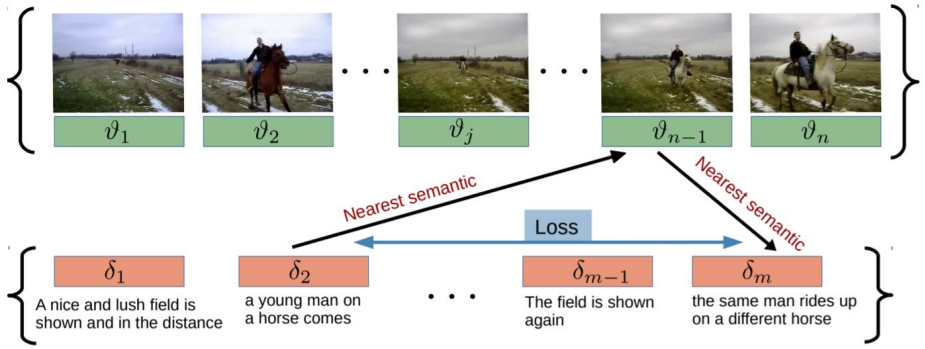


Figure 3: **Cross-Modality Cycle-Consistency.** Starting from a sentence s_i, we find its nearest neighbor in the clip sequence and again its neighbor in the sentence sequence. Deviations from the start index are penalized as alignment error.

Ging et al., COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. NeurIPS 2020.

MERLOT (**M**ultimodal **E**vent **R**epresentation **L**earning **O**ver **T**ime)

- Objectives: i) MLM (mask visual tokens only), ii) VL-NCE (on frames), and iii) temporal reordering (similar to FOM in HERO).
- It combines reconstructive objective and contrastive objective.

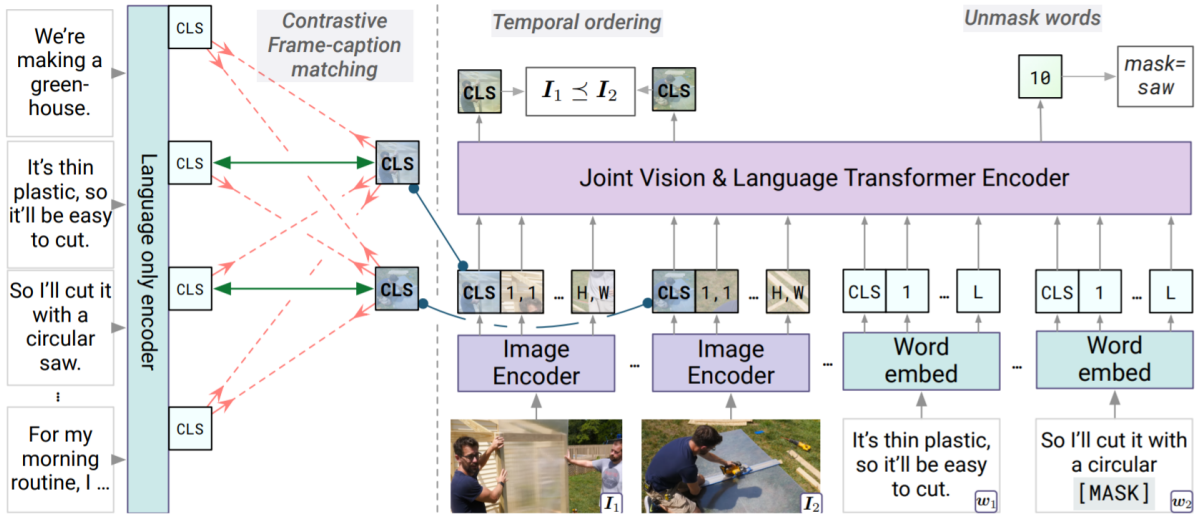
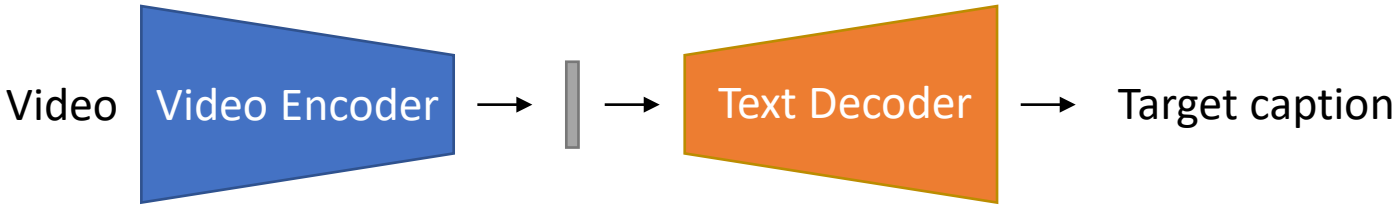


Figure 2: Left: MERLOT learns to match contextualized captions with their corresponding video frames. Right: the same image encoding is provided, along with (masked) word embeddings, into a joint vision-language Transformer model; it then unmasks ground words (like 'saw' in this example) and puts scrambled video frames into the correct order.

Zellers et al., MERLOT: Multimodal Neural Script Knowledge Models. arXiv 2021.

Generative Methods

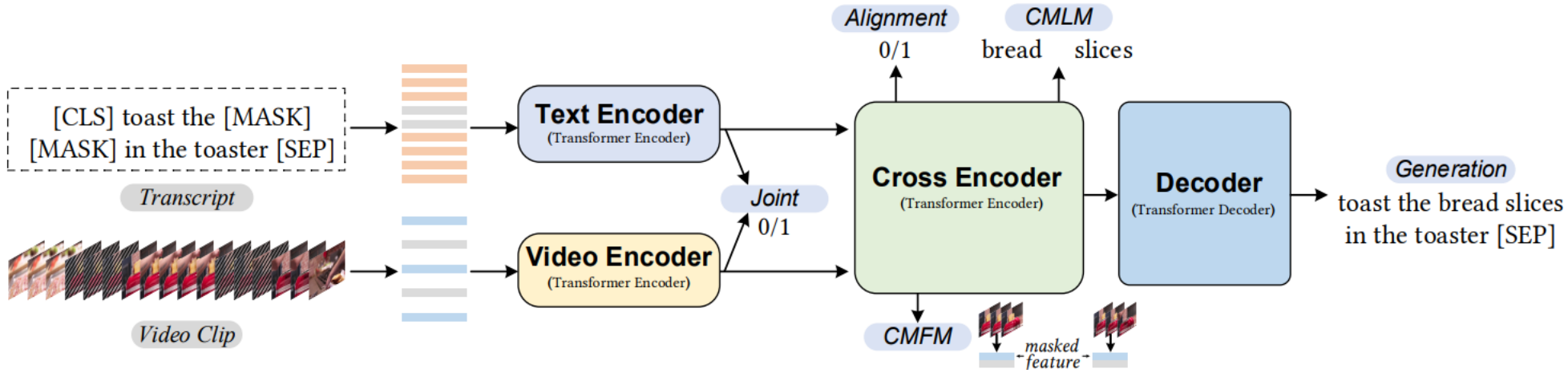
- Video captioning inspired; usually adopt the encoder-decoder architecture



- Leverage video-to-text generation for video representation learning
- Image counterpart: VirTex

UniVL (**Uni**fied **V**ideo and **L**anguage)

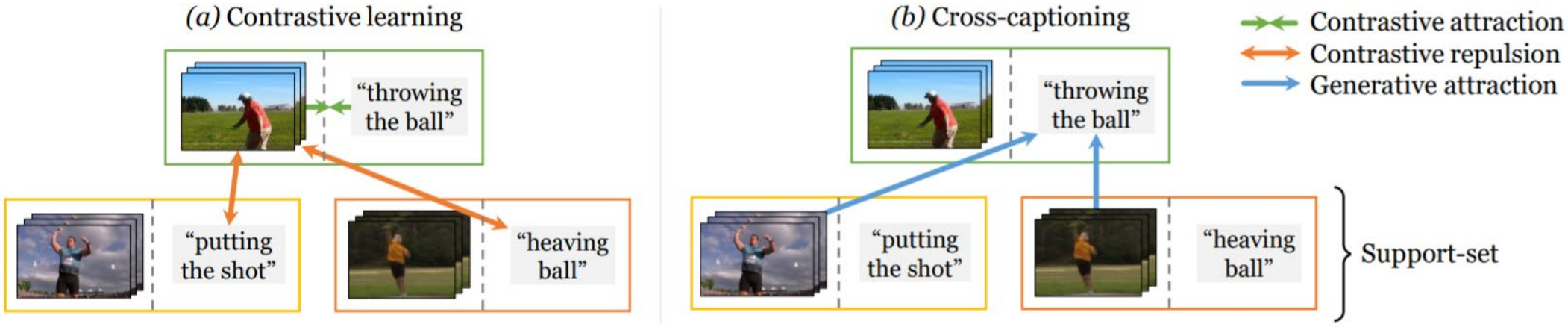
- Objectives: VL-NCE, MLM, MFM, VTM; New: language reconstruction



Luo et al., UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arXiv 2020.

SSB (**S**upport-**S**et **B**ottlenecks)

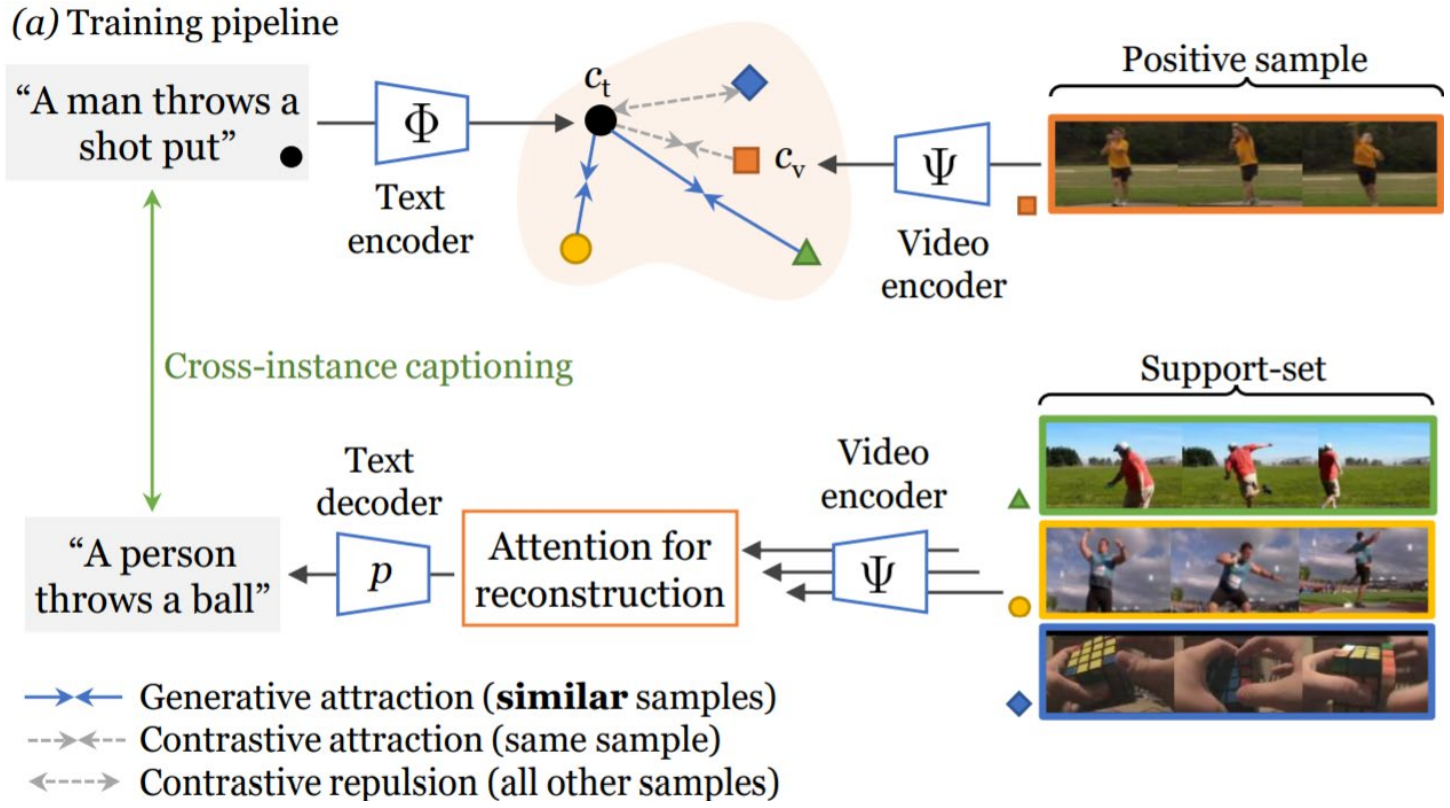
- VL-NCE loss pushes away even semantically related captions.
- This paper introduces cross-captioning, which alleviates this by learning to reconstruct a sample's text representation as a weighted combination of a *support-set*.



Patrick et al., Support-set bottlenecks for video-text representation learning. ICLR 2021.

SSB (**S**upport-**S**et **B**ottlenecks)

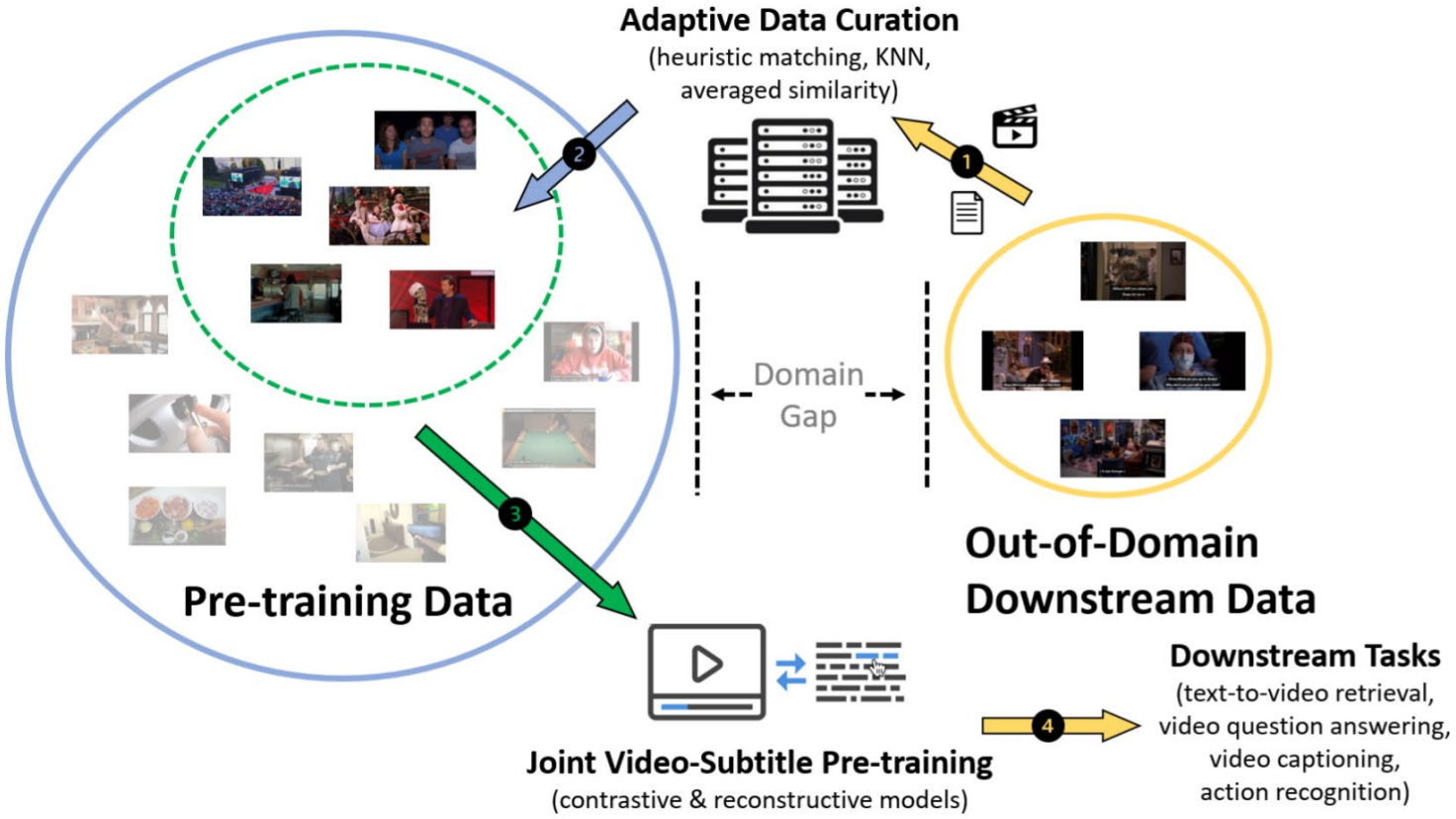
- A *support-set* contains every sample in the minibatch other than the positive sample.



Patrick et al., Support-set bottlenecks for video-text representation learning. ICLR 2021.

CUPID (Adaptive **Cu**ration of **P**re-tra**i**ning **D**ata)

- Close the source-target domain gap



Zhou et al., CUPID: Adaptive Curation of Pre-training Data for Video-and-Language Representation Learning. arXiv 2021.

CUPID (Adaptive **Cu**ration of **P**re-tra**i**ning **D**ata)

- The paradigm is generic and has been applied to various models including MIL-NCE, HERO, CLIP, VLP.

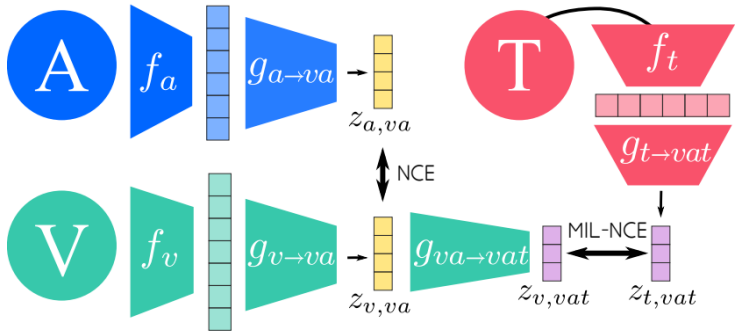
Downstream Task	Domain	Method	Feature Extractor	Our Curated Pre-training
TVR and TVQA	Out-of-domain	HERO	ImageNet/Kinetics pre-trained	200k from HowTo100M and TV
HMDB51	Near-domain	CUPID-CLIP	WIT [47] pre-trained backbone	15k from HowTo100M
YouCook2 captioning	In-domain	CUPID-VLP	HowTo100M pre-trained	15k from HowTo100M
YouCook2 retrieval	In-domain	MIL-NCE	HowTo100M pre-trained backbone	15k from HowTo100M

Table 4. A summary of downstream tasks, domain genres, methods, and pre-training settings.

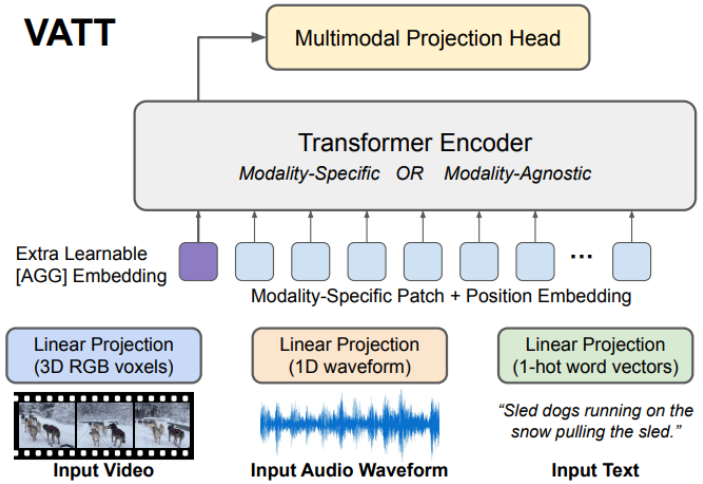
Zhou et al., CUPID: Adaptive Curation of Pre-training Data for Video-and-Language Representation Learning. arXiv 2021.

Other Modalities (Video-Language-Audio)

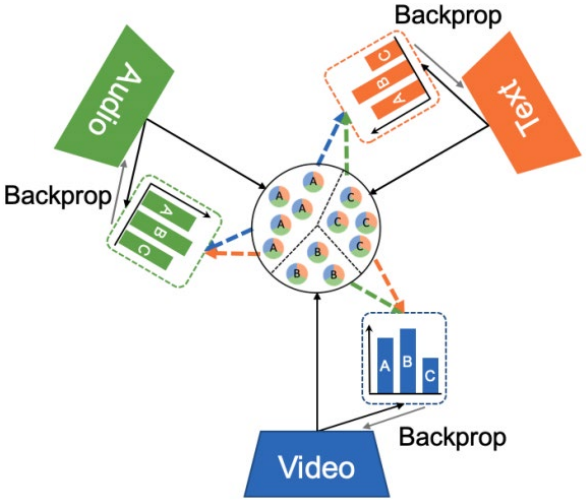
Multi-Modal Versatile Network (MMV)



Video-Audio-Text Transformer (VATT)



Multimodal Clustering Networks (MCN)



Alayrac et al., Self-Supervised MultiModal Versatile Networks. NeurIPS 2020.
Akbari et al., VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. arXiv 2021.
Chen et al., Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos. arXiv 2021.

Other Modalities

- Multimodal Transformer, MMT (ECCV 2020): mixture of seven types/experts of video features, including audio, appearance, motion, speech, scene, face, OCR for overlaid text, for video representation.

- Video-Audio: XDC/GDT/STiCA/AVID etc.

Gabeur et al., Multi-modal Transformer for Video Retrieval. ECCV 2020.

Image-Video Connector

- Can visual representation learned from video pre-training be useful for image tasks?
  - Yes. MMV (NeurIPS 2020) and VATT have results on ImageNet classification. MERLOT have results on VCR (a VQA dataset).

- Joint video-image encoder:

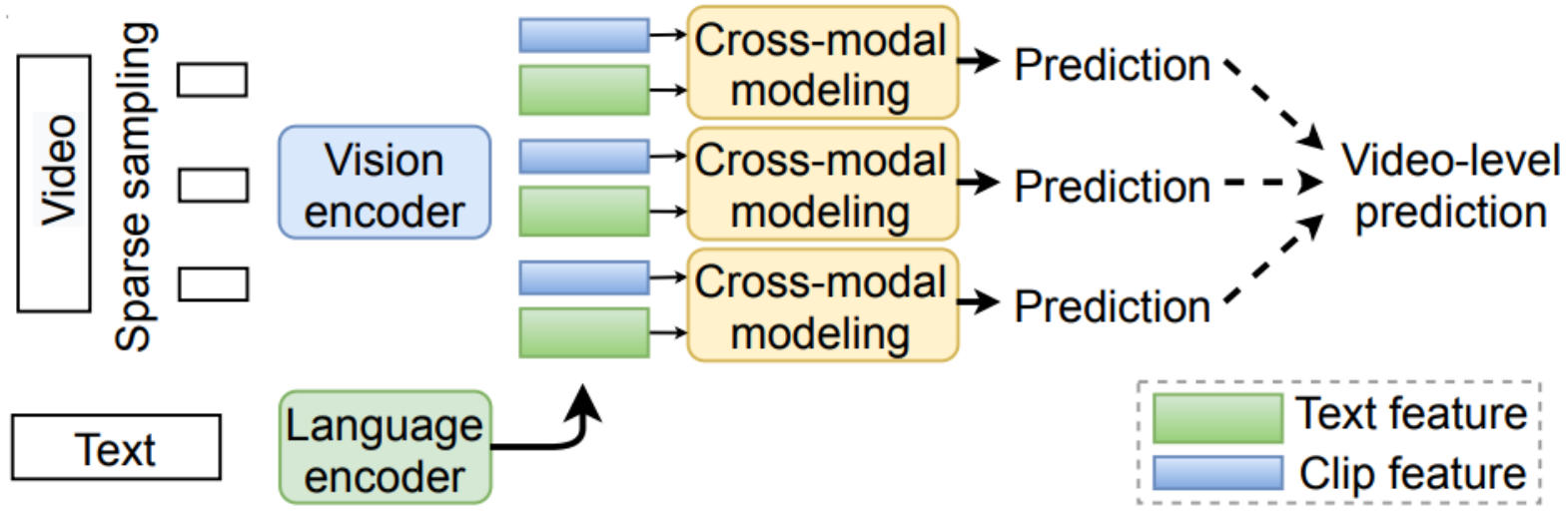
Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval

Max Bain[1] Arsha Nagrani[1†] Gül Varol[1,2] Andrew Zisserman[1]
[1] Visual Geometry Group, University of Oxford
[2] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS
{maxbain, arsha, gul, az}@robots.ox.ac.uk

Image-Video Connector

- On the other hand, can image pre-training benefit video tasks?
  - Yes. See CLIP (OpenAI) and ClipBERT (CVPR 2021 Best Paper Nominee).

- ClipBERT

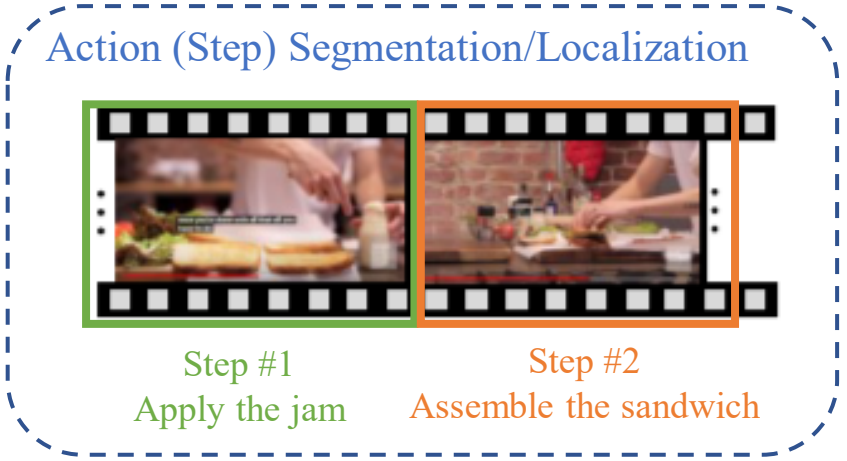


Lei et al., Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling. CVPR 2021.

Method \ Objective	Reconstructive			Contrastive			Generative	Others
	MLM	MFM	FOM	VTM	VL-NCE	Margin	Decoder	
VideoBERT (ICCV 2019)	✓	✓		✓				
ActBERT (CVPR 2020)	✓			✓				Masked Action/Object Classification
HERO (EMNLP 2020)	✓	✓	✓	✓				Video-Subtitle Matching
DECEMBERT (NAACL'21)	✓			✓				Constrained Attention Loss
CBT (arXiv 2019)		✓			✓			
MIL-NCE (CVPR 2020)					✓			MIL version. Same for MMV, VATT.
COOT (NeurIPS 2020)						✓		Cross-modality Cycle-consistency Loss
MERLOT (arXiv 2021)	✓		✓		✓			
UniVL (arXiv 2020)	✓	✓		✓	✓		✓	
SSB (ICLR 2021)					✓		✓	
CUPID (arXiv 2021)	✓	✓	✓	✓	✓		✓	
ClipBERT (CVPR 2021)	✓			✓				

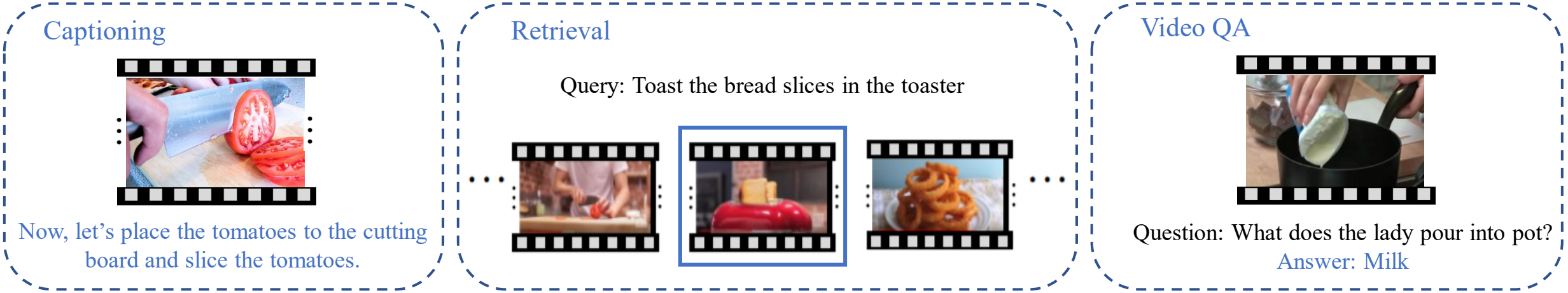
Downstream Tasks and Datasets

- Video-only tasks
  - Action Recognition: **HMDB51, UCF101**, Kinetics-600
  - Action Segmentation/Localization: COIN, CrossTask etc.



Downstream Tasks and Datasets

- Video-Language tasks
  - Video Captioning: **YouCook2**, **MSR-VTT**, VATEX, TVC
  - Text-to-Video Retrieval: **YouCook2**, **MSR-VTT**, DiDeMo, ActivityNet Captions, TVR, VATEX, How2R, MSVD
  - Video QA: **MSRVTT-QA**, TGIF-QA, TVQA, How2QA



Benchmark Results (Video-Only)

- Action Recognition

Method	Modality	Pre-training data	HMDB51	UCF101
Supervised (Duan et al., ECCV 2020)	V	K400+OS	**83.8**	**98.6**
Supervised backbone (SSB, ICLR 2021)	V+T	HowTo+IG65+IM	81.3	98.0
Pure vision-based (Qian et al., CVPR 2021)	V	K600	**70.6**	**94.4**
CBT (arXiv 2019)	V+T	HowTo+ K600	44.5	79.5
MIL-NCE (CVPR 2020)	V+T	HowTo100M	61.0	91.3
MMV (NeurIPS 2020)	V+T+A	HowTo+AudioSet	**75.0**	**95.2**

- Multimodal pre-training has an edge over pure vision-based methods.
- Self-supervised methods are still trailing supervised counterparts.

SSB uses a backbone that is pretrained on IG65M model and another one pretrained on Imagenet. Others are from scratch.

Benchmark Results (Video-Language)

- YouCook2 captioning (video input only)

Method	Pre-training data	BLEU@4	METEOR	CIDEr
Masked Transformer (CVPR 2018)	None	3.85	10.68	37.9
VideoBERT (ICCV 2019)	312K videos	4.33	11.94	55.0
CBT (arXiv 2019)	HowTo+K600	5.12	12.97	64.0
ActBERT (CVPR 2020)	HowTo100M	5.41	13.30	65.0
CUPID (arXiv 2021)	HowTo100M	9.34	16.47	110.5
UniVL (arXiv 2020)	HowTo100M	**11.17**	**17.57**	**127.0**

Pre-training substantially boost performance

Note: results are on *micro-level* metrics. For *macro-level* and *paragraph-level* metrics, see https://github.com/LuoweiZhou/YouCook2-Leaderboard#video-captioning

Benchmark Results (Video-Language)

- YouCook2 text-to-video retrieval (video only, no audio)

Method	Pre-training data	R@1	R@5	R@10	Median R
HGLMM (CVPR 2015)	None	4.6	14.3	21.6	75
Miech et al. (ICCV 2019)	None	4.2	13.7	21.5	65
COOT (NeurIPS 2020)	None	5.9	16.7	24.8	50
Miech et al. (zero-shot)	HowTo100M	6.1	17.3	24.8	46
ActBERT (zero-shot)	HowTo100M	9.6	26.7	38.0	19
MIL-NCE (zero-shot)	HowTo100M	13.9	36.3	48.9	11
MMV (zero-shot)	HowTo+AudioSet	11.7	33.4	45.4	13
MCN (zero-shot)	HowTo100M	18.1	35.5	45.2	-
Miech et al. (ICCV 2019)	HowTo100M	8.2	24.5	35.3	24
COOT (NeurIPS 2020)	HowTo100M	16.7	40.2	52.3	9
DECEMBERT (NAACL 2021)	HowTo100M	17.0	**43.8**	**59.8**	9
CUPID (arXiv 2021)	HowTo100M	**17.7**	43.2	57.1	**7**
UniVL (arXiv 2020)	HowTo100M	**28.9**	**57.6**	**70.0**	**4**

Pre-trained models generalize well

Pre-training wins again

Benchmark Results (Video-Language)

- MSR-VTT text-to-video retrieval (video only, no audio)

Method	Pre-training data	R@1	R@5	R@10	Median R
SSB, w/o pre-training (ICLR 2021)	None	27.4	56.3	67.7	3
Miech et al. (ICCV 2019)	HowTo100M	14.9	40.2	52.8	9
ActBERT (CVPR 2020)	HowTo100M	16.3	42.8	56.9	10
HERO (EMNLP 2020)	HowTo100M+TV	16.8	43.4	57.7	-
UniVL (arXiv 2020)	HowTo100M	21.2	49.6	63.1	6
NoiseEstimation (AAAI 2021)	HowTo100M	17.4	41.6	53.6	8
SSB (ICLR 2021)	HowTo100M	**30.1**	**58.5**	**69.3**	**3**
ClipBERT (CVPR 2021)	COCO and VG	22.0	46.8	59.9	6
DECEMBERT (NAACL 2021)	HowTo100M	17.5	44.3	58.6	9

Limited gain possibly due to domain discrepancy

Benchmark Results (Video-Language)

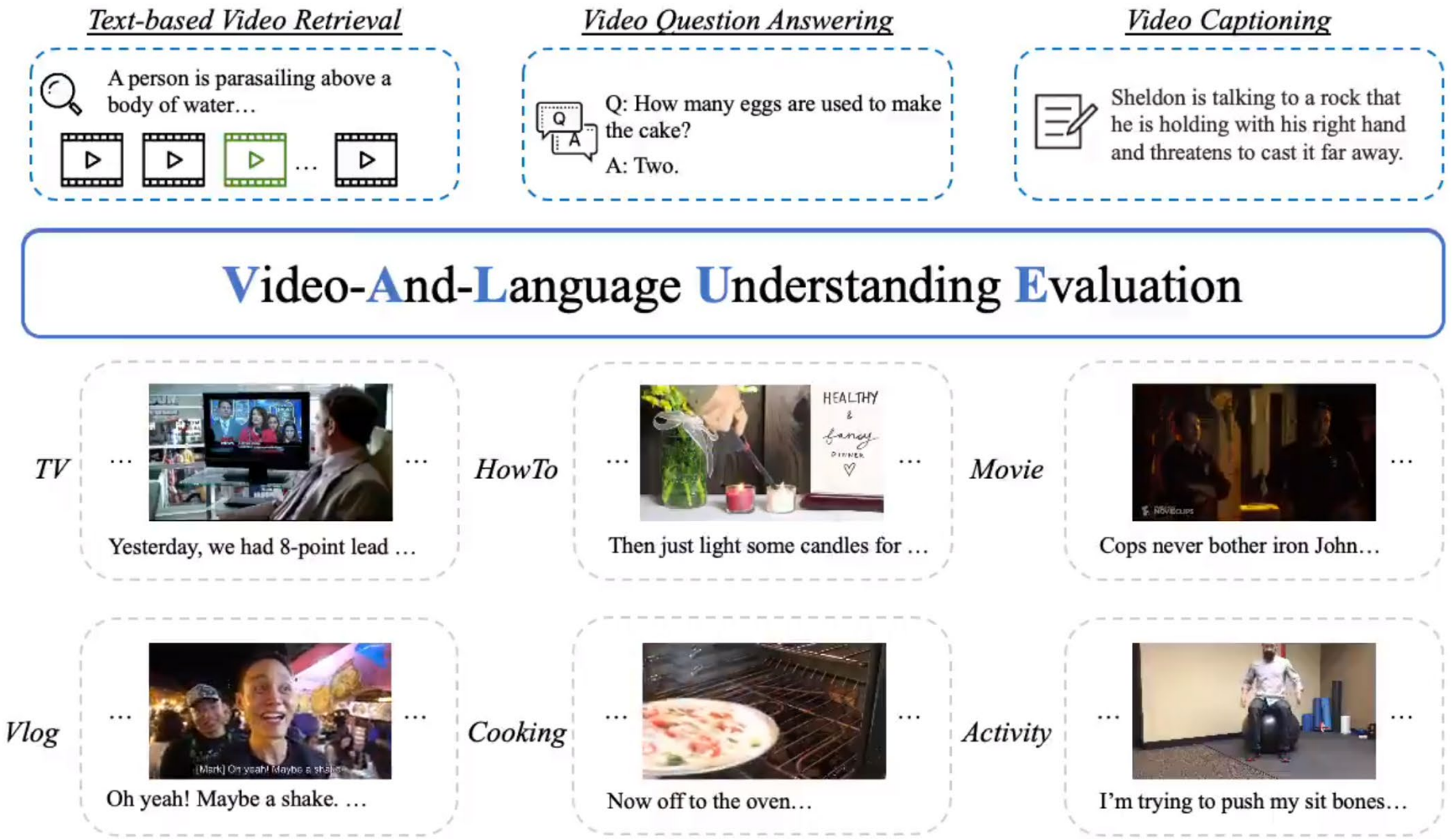
- Video QA

Method	Pre-training data	MSRVTT-QA	TVQA
STAGE (ACL 2020)	None	-	70.23
HCRN (CVPR 2020)	None	27.4	-
HERO (EMNLP 2020)	HowTo100M+TV	-	73.61
NoiseEstimation (AAAI 2021)	HowTo100M	35.1	-
DECEMBERT (NAACL 2021)	HowTo100M	37.4	-
ClipBERT (CVPR 2021)	COCO+VG	37.4	-
CoMVT (CVPR 2021)	HowTo100M	39.5	-
VQA-T (arXiv 2021)	HowToVQA69M	41.5	-
MERLOT (arXiv 2021)	YT-Temporal-180M	**43.1**	**78.7**

YT-Temporal-180M is larger than HowTo100M and contains diverse topics; this allows it to go beyond literal descriptions and capture more commonsense knowledge that could benefit QA.

Seo et al., Look Before you Speak: Visually Contextualized Utterances. CVPR 2021.
Yang et al., Just Ask: Learning to Answer Questions from Millions of Narrated Videos. arXiv 2021.

Video-And-Language Understanding Evaluation (VALUE)



Li et al., VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. arXiv 2021.

Task Name	Video Source	More info	Metric
		Retrieval Tasks	
TVR	TV episodes		Average(R@1, 5, 10) with tIoU >= 0.7
How2R	YouTube (HowTo100M)		Average(R@1, 5, 10) with tIoU >= 0.7
YC2R	YouTube		Average(R@1, 5, 10)
VATEX-EN-R	YouTube		Average(R@1, 5, 10)
		QA Tasks	
TVQA	TV episodes		Accuracy
How2QA	YouTube (HowTo100M)		Accuracy
VIOLIN	TV episodes, Movie clips		Accuracy
VLEP	TV episodes, YouTube		Accuracy
		Captioning Tasks	
TVC	TV episodes		CIDEr-D
YC2C	YouTube		CIDEr-D
VATEX-EN-C	YouTube		CIDEr-D

VALUE competition will be held in conjunction with CLVL workshop at ICCV 2021!

Conclusion

- Video-and-Language Pre-training is a nascent field with great potential.

- Limitations
  - The use of different modalities (video, audio), pretraining datasets (HowTo100M, Kinetics-600), architectures (S3D, SlowFast), pre-training (supervised, self-supervised) makes it difficult to have fair comparisons.
  - More unified benchmarks need to be proposed. VALUE is a good start.

- Future Directions
  - Further scale up the data and its domain diversity
  - Multimodal and multilingual

Thank you!
Any questions?

VALUE Leaderboard