Video-Text Pre-training

Kevin Lin, Linjie Li, Chung-Ching Lin

6/19/2022

Pretrain-and-finetune paradigm

- Pre-train on large amounts of datasets is very helpful for performance improvement on target tasks with small datasets

Object Detection

Instance Segmentation

Semantic Segmentation

…

Vision

BERT (Devlin et al, 2018)

Wikipedia

BookCorpus

Question Answering

Natural Language Inference

Sentence Classification

…

Language

COCO, CC3M, CC12M, CLIP400M, …

Visual Question Answering

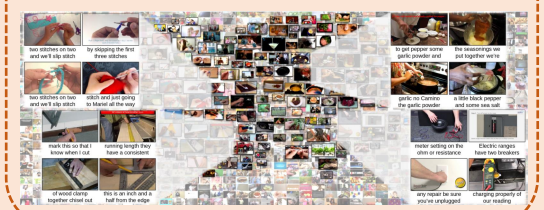
Image/Text Retrieval

Image Captioning

…

Image & Language

HT100M, MERLOT, …



Video Question Answering

Video-Text Retrieval

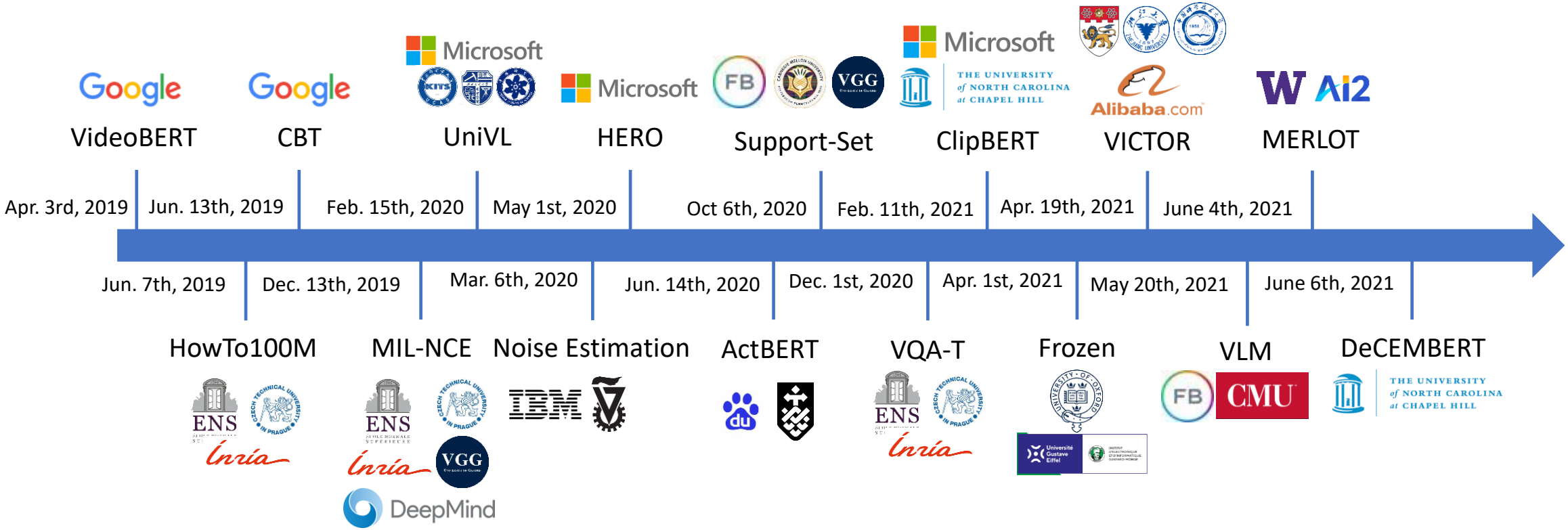
Video Captioning

…

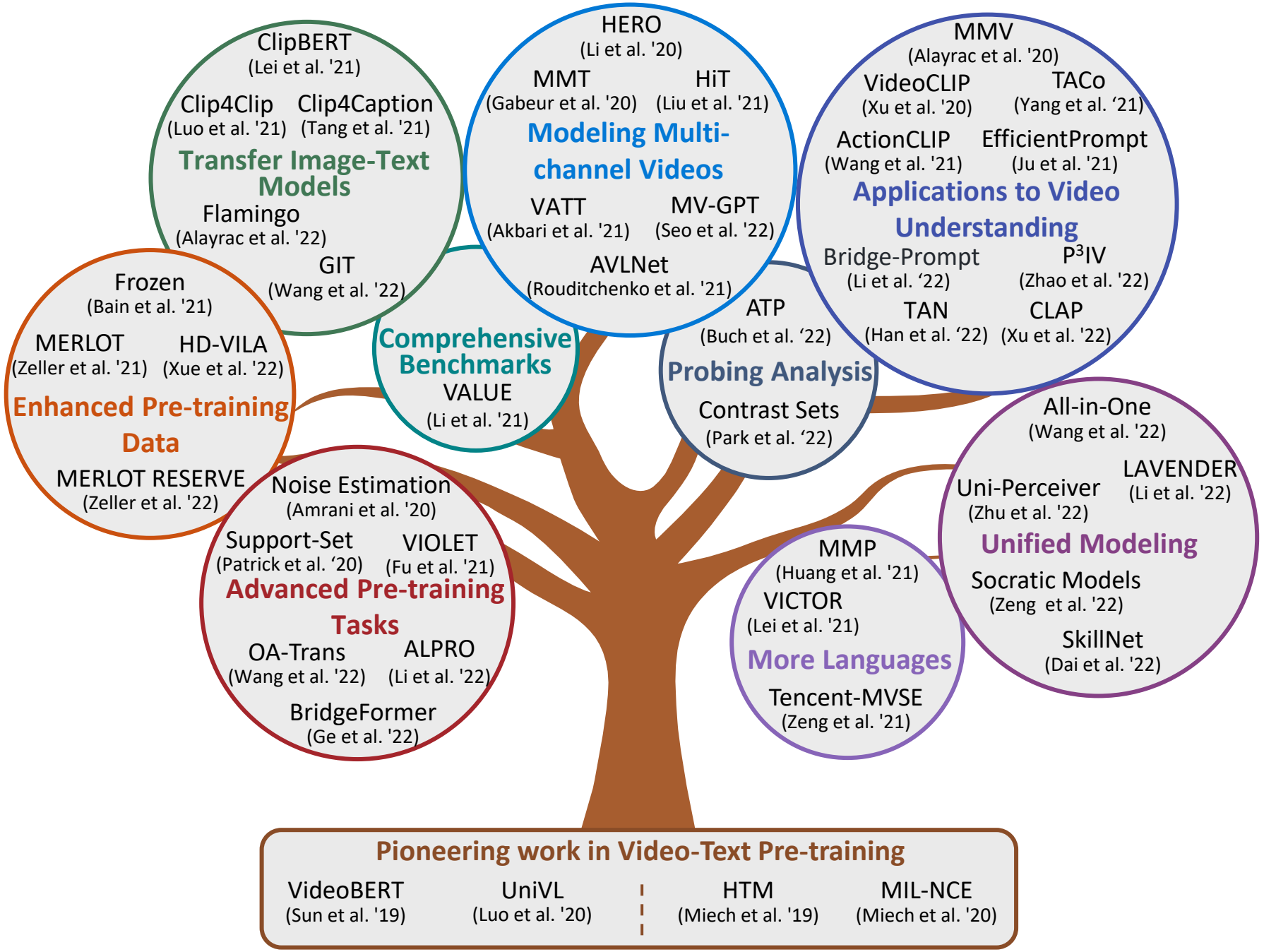
Video & Language

Evolution of Video-Text Pre-training

Representative Video-Text Models until CVPR 2021



Many more methods have been proposed since then …


ClipBERT
(Lei et al. '21)
Clip4Clip
(Luo et al. '21)
Clip4Caption
(Tang et al. '21)
Transfer Image-Text Models
Flamingo
(Alayrac et al. '22)
GIT
(Wang et al. '22)
HERO
(Li et al. '20)
MMT
(Gabeur et al. '20)
HiT
(Liu et al. '21)
Modeling Multi-channel Videos
VATT
(Akbari et al. '21)
MV-GPT
(Seo et al. '22)
AVLNet
(Rouditchenko et al. '21)
MMV
(Alayrac et al. '20)
VideoCLIP
(Xu et al. '20)
TACo
(Yang et al. '21)
ActionCLIP
(Wang et al. '21)
EfficientPrompt
(Ju et al. '21)
Applications to Video Understanding
Bridge-Prompt
(Li et al. '22)
P3IV
(Zhao et al. '22)
TAN
(Han et al. '22)
CLAP
(Xu et al. '22)
Frozen
(Bain et al. '21)
MERLOT
(Zeller et al. '21)
HD-VILA
(Xue et al. '22)
Enhanced Pre-training Data
MERLOT RESERVE
(Zeller et al. '22)
Comprehensive Benchmarks
VALUE
(Li et al. '21)
ATP
(Buch et al. '22)
Probing Analysis
Contrast Sets
(Park et al. '22)
All-in-One
(Wang et al. '22)
Uni-Perceiver
(Zhu et al. '22)
LAVENDER
(Li et al. '22)
Unified Modeling
Socratic Models
(Zeng et al. '22)
SkillNet
(Dai et al. '22)
Noise Estimation
(Amrani et al. '20)
Support-Set
(Patrick et al. '20)
VIOLET
(Fu et al. '21)
Advanced Pre-training Tasks
OA-Trans
(Wang et al. '22)
ALPRO
(Li et al. '22)
BridgeFormer
(Ge et al. '22)
MMP
(Huang et al. '21)
VICTOR
(Lei et al. '21)
More Languages
Tencent-MVSE
(Zeng et al. '21)
Pioneering work in Video-Text Pre-training
VideoBERT
(Sun et al. '19)
UniVL
(Luo et al. '20)
HTM
(Miech et al. '19)
MIL-NCE
(Miech et al. '20)

Agenda

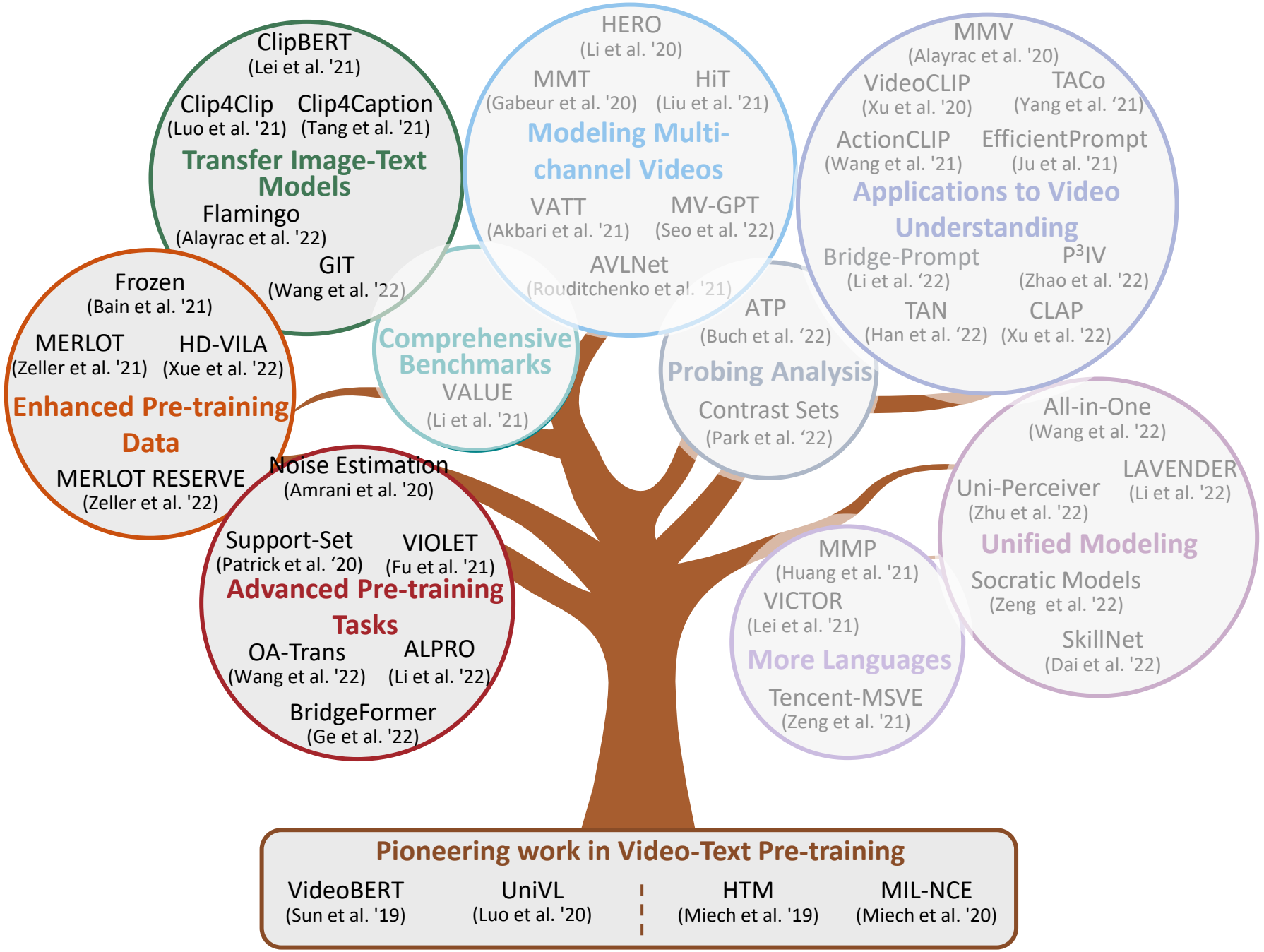
Overview of Video-Text Pre-training

Learning from Multi-channel Videos: Methods and Benchmarks

Advanced Topics in Video-Text Pretraining

Overview of Video-Text Pre-training

Kevin Lin

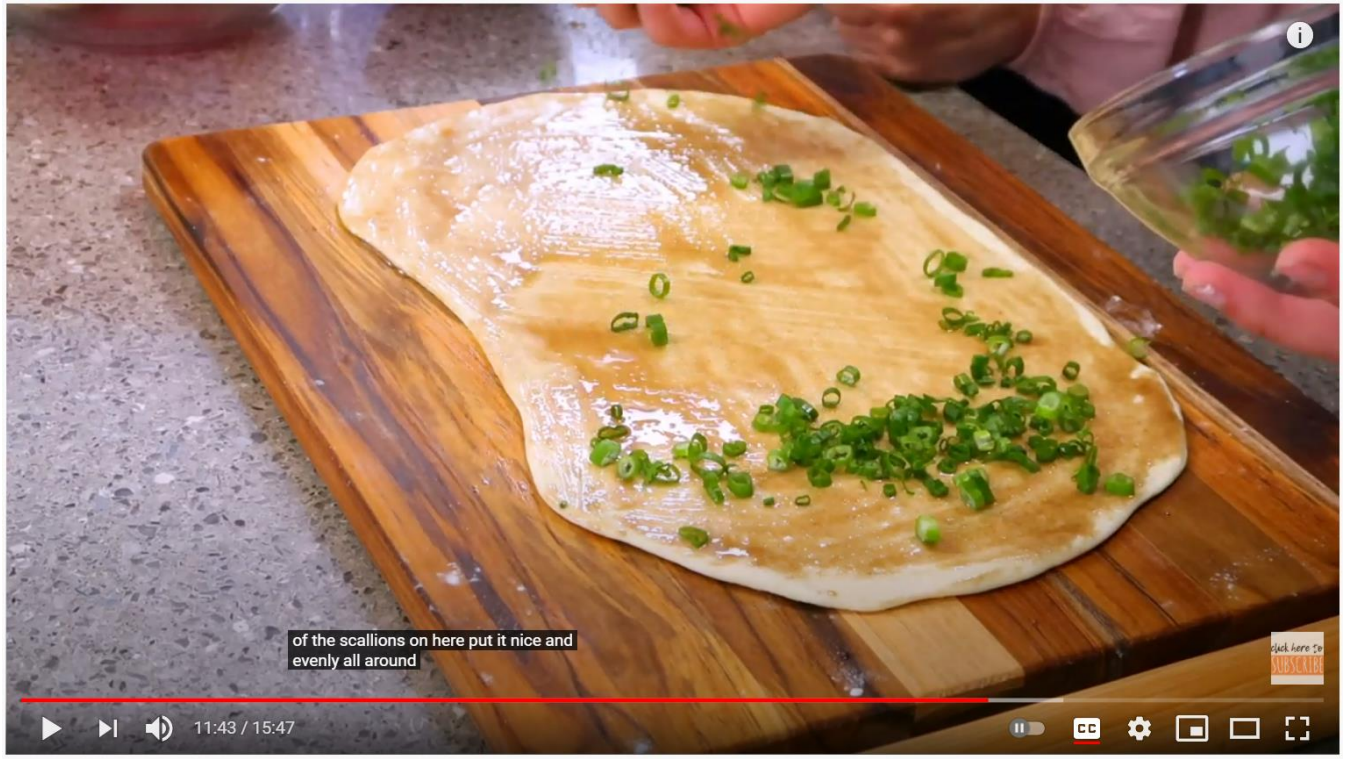

ClipBERT
(Lei et al. '21)
Clip4Clip
(Luo et al. '21)
Clip4Caption
(Tang et al. '21)
Transfer Image-Text Models
Flamingo
(Alayrac et al. '22)
GIT
(Wang et al. '22)
HERO
(Li et al. '20)
MMT
(Gabeur et al. '20)
HiT
(Liu et al. '21)
Modeling Multi-channel Videos
VATT
(Akbari et al. '21)
MV-GPT
(Seo et al. '22)
AVLNet
(Rouditchenko et al. '21)
MMV
(Alayrac et al. '20)
VideoCLIP
(Xu et al. '20)
TACo
(Yang et al. '21)
ActionCLIP
(Wang et al. '21)
EfficientPrompt
(Ju et al. '21)
Applications to Video Understanding
Bridge-Prompt
(Li et al. '22)
P3IV
(Zhao et al. '22)
TAN
(Han et al. '22)
CLAP
(Xu et al. '22)
Frozen
(Bain et al. '21)
MERLOT
(Zeller et al. '21)
HD-VILA
(Xue et al. '22)
Enhanced Pre-training Data
MERLOT RESERVE
(Zeller et al. '22)
Comprehensive Benchmarks
VALUE
(Li et al. '21)
ATP
(Buch et al. '22)
Probing Analysis
Contrast Sets
(Park et al. '22)
All-in-One
(Wang et al. '22)
Uni-Perceiver
(Zhu et al. '22)
LAVENDER
(Li et al. '22)
Unified Modeling
Socratic Models
(Zeng et al. '22)
SkillNet
(Dai et al. '22)
Noise Estimation
(Amrani et al. '20)
Support-Set
(Patrick et al. '20)
VIOLET
(Fu et al. '21)
Advanced Pre-training Tasks
OA-Trans
(Wang et al. '22)
ALPRO
(Li et al. '22)
BridgeFormer
(Ge et al. '22)
MMP
(Huang et al. '21)
VICTOR
(Lei et al. '21)
More Languages
Tencent-MSVE
(Zeng et al. '21)
Pioneering work in Video-Text Pre-training
VideoBERT
(Sun et al. '19)
UniVL
(Luo et al. '20)
HTM
(Miech et al. '19)
MIL-NCE
(Miech et al. '20)

Outline

- Data and challenges
- Pioneer work in video-text pre-training
- Advanced pre-training tasks
- Transferring image-text model

Video-and-Language Pre-training

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



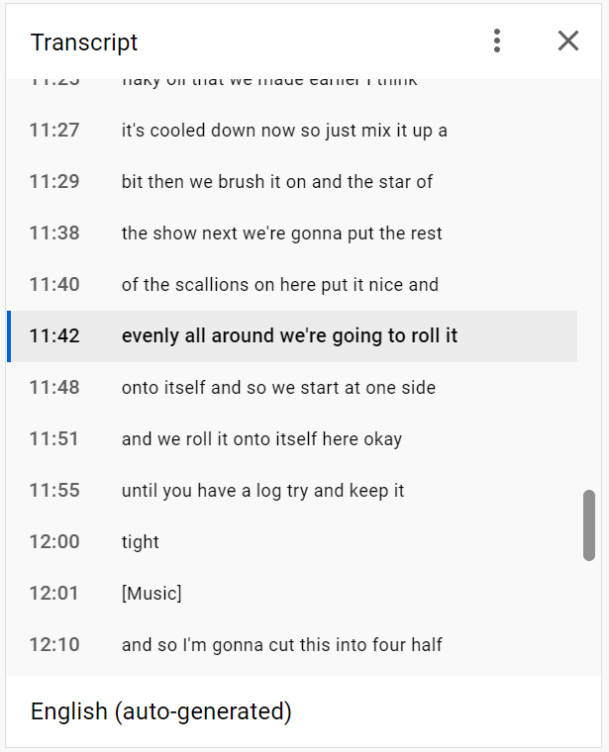


Figure credit: Making Scallion Pancake Beef Rolls: https://www.youtube.com/watch?v=vTmgLKtx49Y
Slide credit: CVPR 2021 VQA2VLN Tutorial

Video-and-Language Pre-training

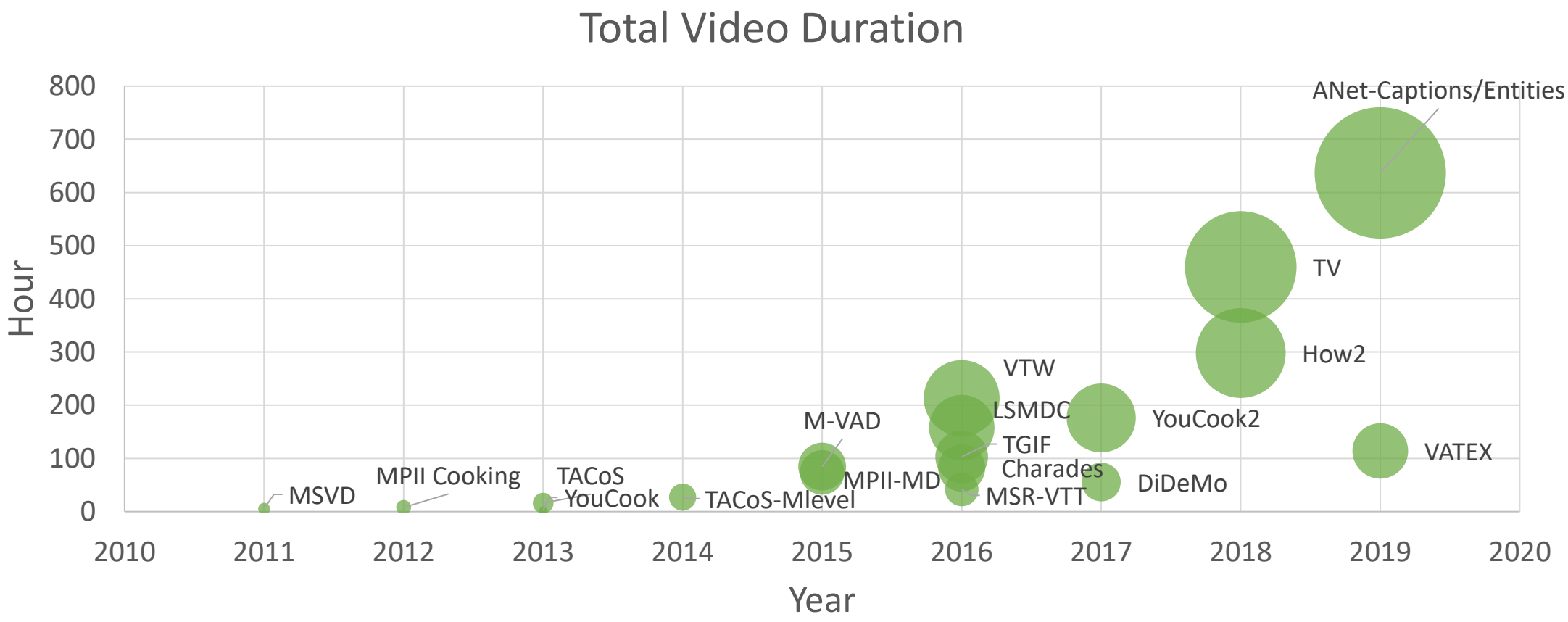
- Paired video clips and subtitles

"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.

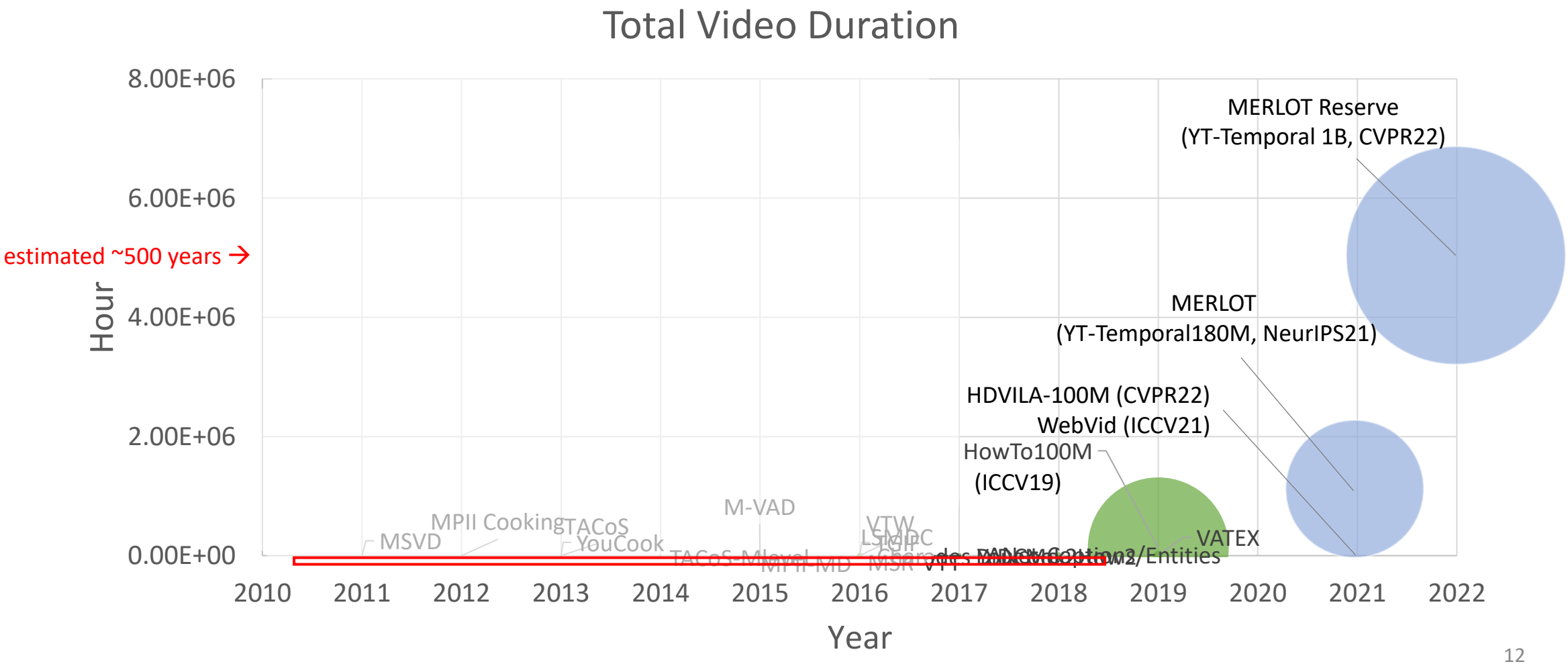
Figure credit: https://ai.googleblog.com/2019/09/learning-cross-modal-temporal.html
Slide credit: CVPR 2021 VQA2VLN Tutorial

Evolution of Video-and-Language Datasets



Credit: CVPR 2021 VQA2VLN Tutorial

Evolution of Video-and-Language Datasets



Merlot: Multimodal neural script knowledge models, NeurIPS 2021

HowTo100M

- 136M video clips from YouTube videos
- Each clip is paired with an automatically transcribed narration
- 23K activities

Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades [48]	10k	16k	10,000	82h	Home	2016
MSR-VTT [58]	10k	200k	7,180	40h	Youtube	2016
YouCook2 [67]	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS [7]	40k	40k	432	55h	Home	2018
DiDeMo [15]	27k	41k	10,464	87h	Flickr	2017
M-VAD [52]	49k	56k	92	84h	Movies	2015
MPII-MD [43]	69k	68k	94	41h	Movies	2015
ANet Captions [26]	100k	100k	20,000	849h	Youtube	2017
TGIF [27]	102k	126k	102,068	103h	Tumblr	2016
LSMDC [44]	128k	128k	200	150h	Movies	2017
How2 [45]	185k	185k	13,168	298h	Youtube	2018
HowTo100M	**136M**	**136M**	**1.221M**	**134,472h**	Youtube	2019

15 years!

Category	Tasks	Videos	Clips
Food and Entertaining	11504	497k	54.4M
Home and Garden	5068	270k	29.5M
Hobbies and Crafts	4273	251k	29.8M
Cars & Other Vehicles	810	68k	7.8M
Pets and Animals	552	31k	3.5M
Holidays and Traditions	411	27k	3.0M
Personal Care and Style	181	16k	1.6M
Sports and Fitness	205	16k	2.0M
Health	172	15k	1.7M
Education and Communications	239	15k	1.6M
Arts and Entertainment	138	10k	1.2M
Computers and Electronics	58	5k	0.6M
Total	23.6k	1.22M	136.6M

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, ICCV 2019

Challenges in training data

- Noisy transcript (automatically generated with ASR tools)
- Constrained domains (instruction videos)
- Temporally misaligned
- Computing resources demanding

End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

Video Transcription vs. Caption

Transcript (Casual speech, fragmentary, lacking punctuation, etc)

now I'm just kind of grilling these tomatoes in this pan I want to get the maximum flavor usually you always use tomatoes raw as it but I just want to add that little dimension of cooked a slightly charged tomatoes yum!

Caption (Formal and concise)

Grill the tomatoes in a pan and then put them on a plate.

Language styles are different!

Examples from Youcook dataset: Towards Automatic Learning of Procedures From Web Instructional Videos, AAAI 2018

Recent Data Sources

- Researchers have been working on collecting better quality (well-aligned) data for the pre-training.

Smiling Beautiful Family of Four Play Catch with Happy Golden Retriever Dog on the Backyard Lawn. Idyllic Family Has Fun with Loyal Pedigree Dog Outdoors in Summer House. Slow Motion Shot

YouTube

shutterstock

dreamstime

reddit

> 60 sec

15~60 sec

< 15 sec

Example from shutterstock. Link: [Smiling Beautiful Family Four Play Catch Stock Footage Video (100% Royalty-free) 1058262028 | Shutterstock]

WebVid-2M & WebVid-10M

- Well-aligned video-text pairs from high-quality video sources

Billiards, concentrated young woman playing in club.

Female cop talking on walkietalkie, responding emergency call, crime prevention

Runners feet in a sneakers close up. realistic three dimensional animation.

Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing

dataset	domain	#clips	avg dur. (secs)	#sent	time (hrs)
MPII Cook [47]	cooking	44	600	6K	8
TACos [44]	cooking	7K	360	18K	15.9
DideMo [3]	flickr	27K	28	41K	87
MSR-VTT [65]	youtube	10K	15	200K	40
Charades [53]	home	10K	30	16K	82
LSMDC15 [46]	movies	118K	4.8	118K	158
YouCook II [70]	cooking	14K	316	14K	176
ActivityNet [24]	youtube	100K	180	100K	849
CMD [5]	movies	34K	132	34K	1.3K
WebVid-2M	open	**2.5M**	18	**2.5M**	**13K**
HT100M [37]	instruction	136M	4	136M	134.5K

Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval, ICCV 2021

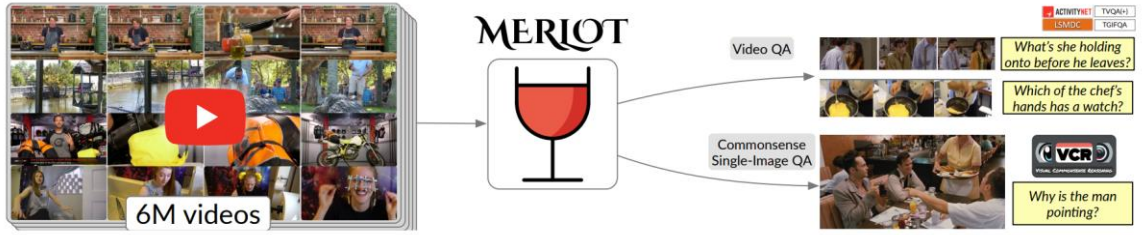
Scale It Up

- Collect larger scale, more diverse videos from YouTube

YT-Temporal 180M (NeurIPS21):
30% of videos are about local news & monetized contents

YT-Temporal 1B (CVPR22):
Scale it up in terms of video domains and # videos

HD-VILA-100M (CVPR22):
High-resolution videos (720p)



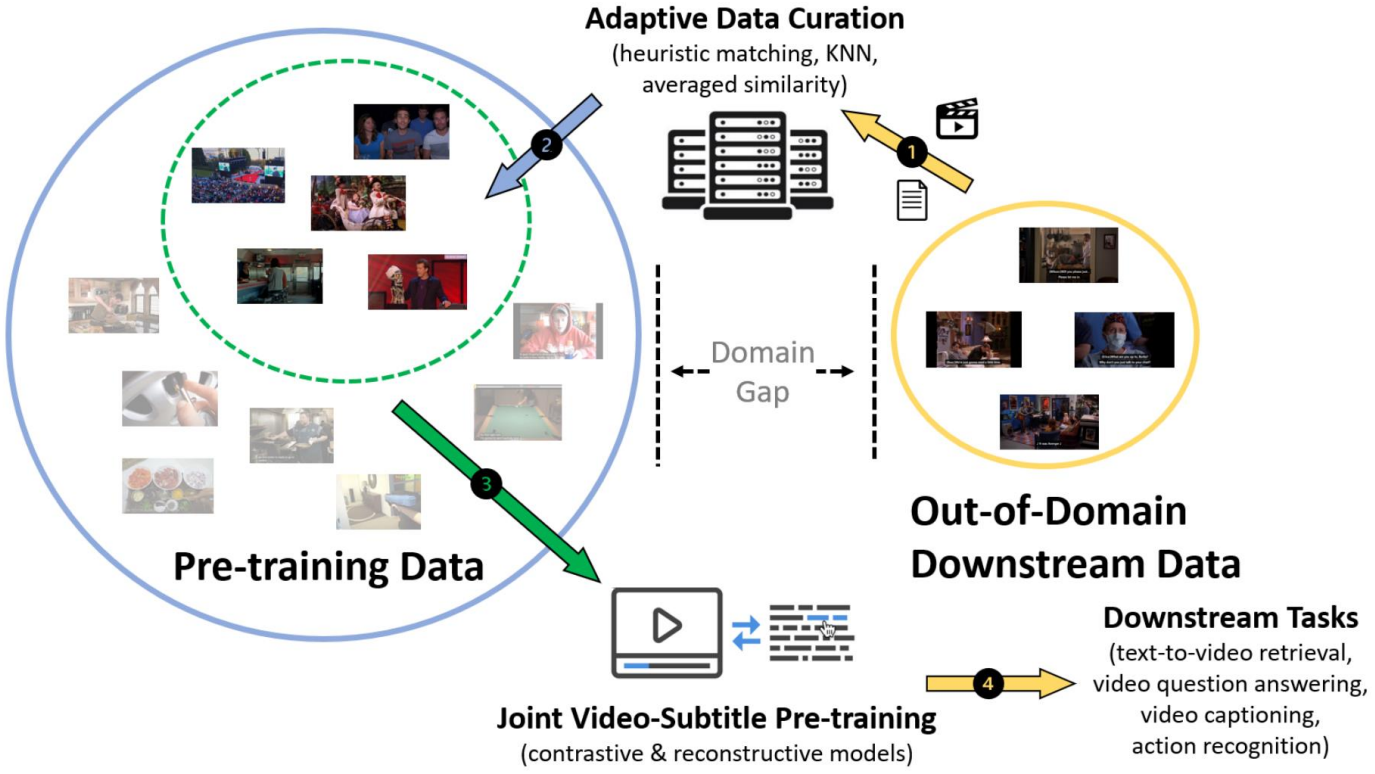
Dataset	Domain	#Video clips	#Sentence	Avg len(sec)	Sent len	Duration(h)	Resolution
MSR-VTT [53]	open	10K	200K	15.0	9.3	40	240p
DideMo [2]	Flickr	27K	41K	6.9	8.0	87	-
LSMDC [41]	movie	118K	118K	4.8	7.0	158	1080p
YouCook II [62]	cooking	14K	14K	19.6	8.8	176	-
How2 [43]	instructional	80K	80K	90.0	20.0	2K	-
ActivityNet Caption [25]	action	100K	100K	36.0	13.5	849	-
WebVid-2M [3]	open	2.5M	2.5M	18.0	12.0	13K	360p
HowTo100M [37]	instructional	136M	136M	3.6	4.0	134.5K	240p
HD-VILA-100M (Ours)	open	103M	103M	13.4	32.5	371.5K	720p

Table 1. Statistics of HD-VILA-100M and its comparison with existing video-language datasets.

MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound, CVPR 2022
Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions, CVPR 2022

The more the better?

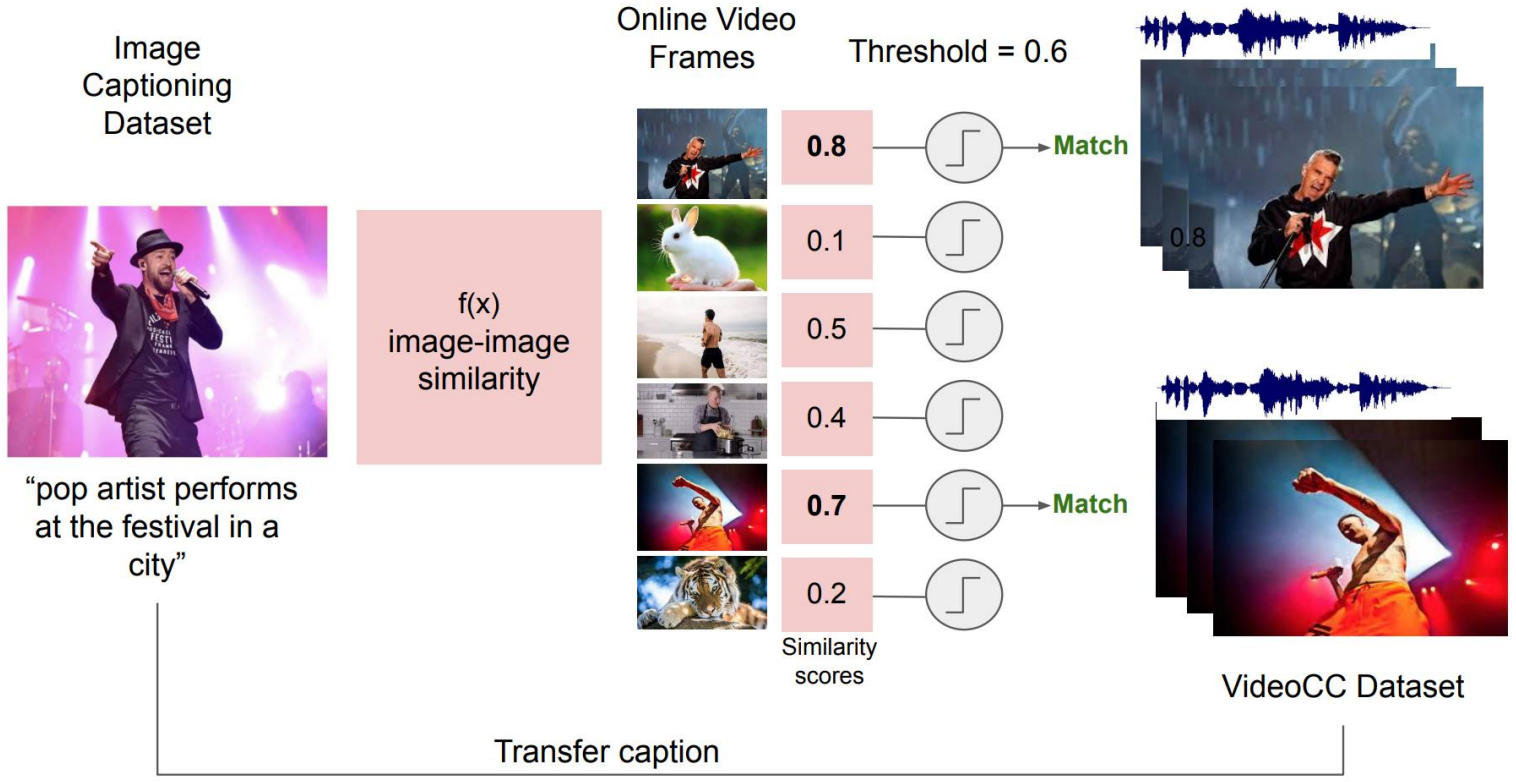
- Researchers are exploring to use a small subset of data for domain-specific pre-training



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

The more the better?

- VideoCC3M: Mining audio-video clips



Pretraining Data	Modality	# Caps	R@1	R@5	R@10
Finetuned					
-	V	-	30.2	60.7	71.1
HowTo100M [55]	V	130M	33.1	62.3	72.3
VideoCC3M	V	970K	35.0	63.1	75.1
VideoCC3M	A+V	**970K**	**35.8**	**65.1**	**76.9**
Zero-shot					
HowTo100M [55]	V	130M	8.6	16.9	25.8
VideoCC3M	V	970K	18.9	37.5	47.1
VideoCC3M	A+V	**970K**	**19.4**	**39.5**	**50.3**

Table 2. **Effect of pretraining data on text-video retrieval for the MSR-VTT dataset. # Caps:** Number of unique captions. Training on VideoCC3M provides much better performance than Howto100M, with a fraction of the dataset size (VideoCC3M has only 970K captions and 6.3M clips compared to the 130M clips in HowTo100M) . The performance boost is particularly large for the zero-shot setting.

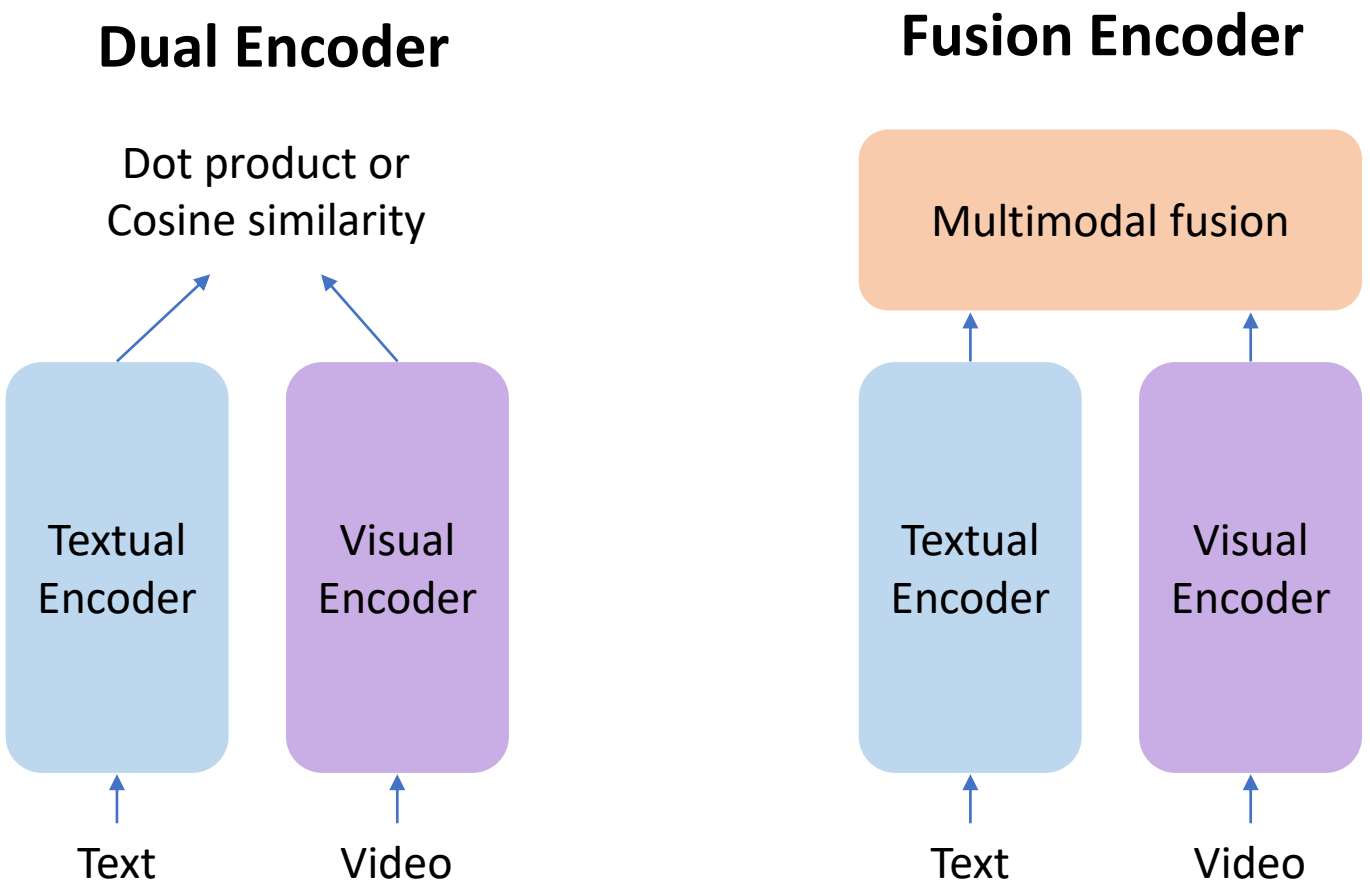
Learning Audio-Video Modalities from Image Captions, arXiv 2022

Outline

- Data and challenges
- Pioneer work in video-text pre-training
- Advanced pre-training tasks
- Transferring image-text model

Model Architecture

Most existing approaches can be roughly classified into two categories

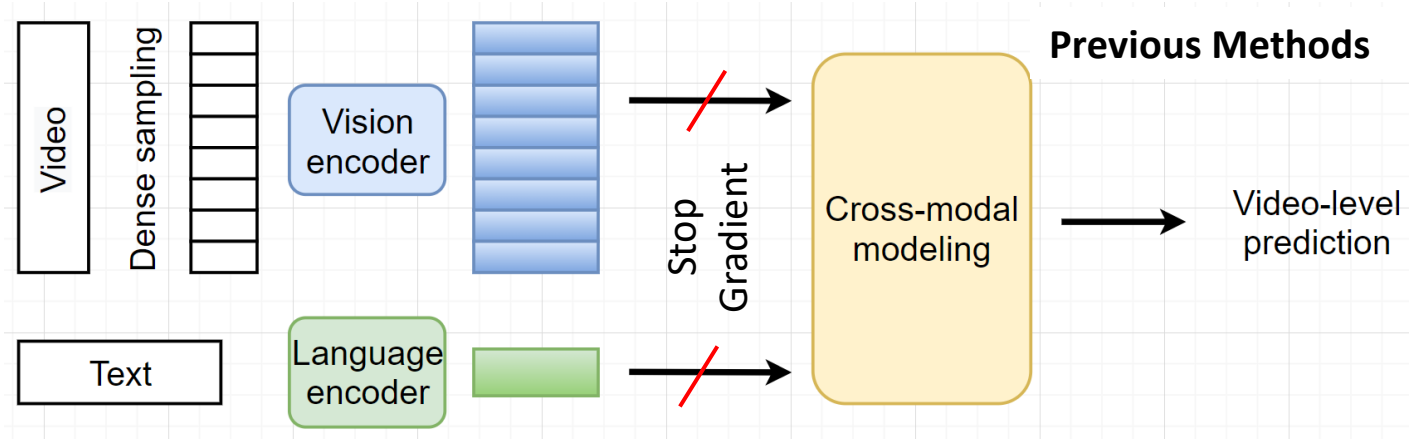


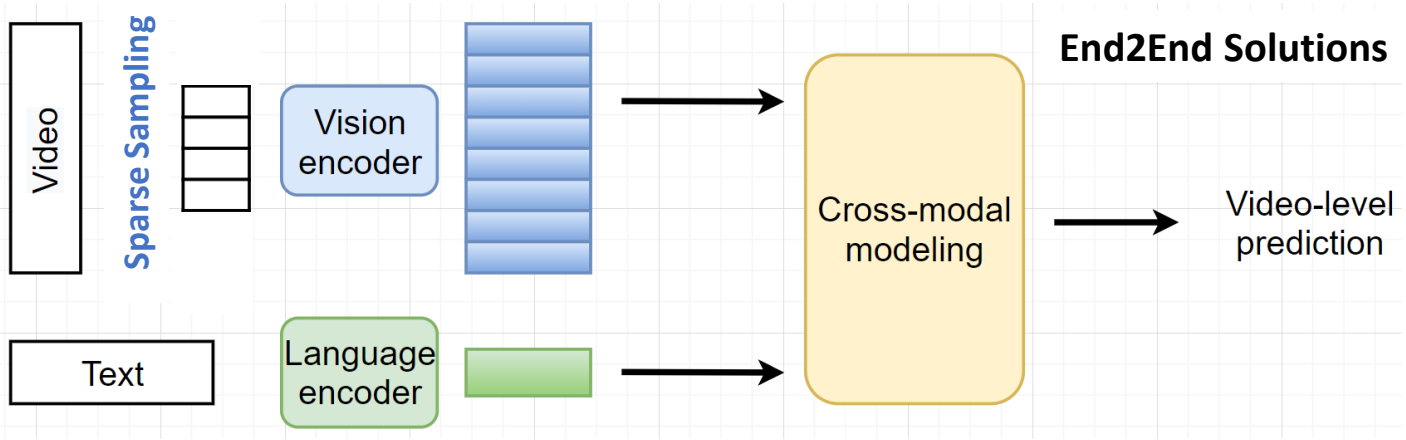
Dual Encoder

- Large-scale contrastive video-text learning
- Favorable architecture for image-text retrieval

[HTM, Miech et al., 2019], [MIL-NCE, Miech et al., 2020], [Support Set, Patrick et al., 2020], [Frozen, Bain et al., 2021], [VideoCLIP, Xu et al., 2021]
Figure credit: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, ICCV 2019

Fusion Encoder





- Deep fusion: better model the interactions between video and text
- Strong improvements on Video QA and Video Captioning

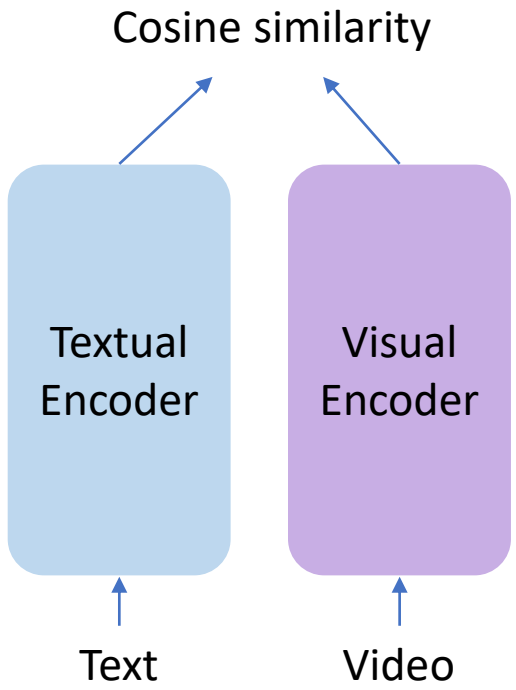
[VideoBERT, Sun et al., 2019], [UniVL, Luo et al., 2020], [ClipBERT, Lei et al.,2021], [MERLOT, Zellers et al.,2021]
Figure credit: Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling, CVPR 2021

Overview of the representative VLP models for video-and-language

Model	Multimodal Fusion	Vision Encoder	Text Encoder	Decoder	E2E	Pre-training Objectives
VideoBERT (Sun et al., 2019b)	Xformer	3D CNN	Emb.	✗	✗	MLM+VTM+MVM
ActBERT (Zhu and Yang, 2020)		OD	Emb.	✗	✗	MLM+VTM+MVM
CBT (Sun et al., 2019a)		3D CNN	Xformer	✗	✗	VTC
HERO (Li et al., 2020b)		2D+3D CNN	Emb.	✗	✗	MLM+VTM+FOM +MFM
UniVL (Luo et al., 2020)		2D+3D CNN +Xformer	Xformer	✓	✗	VTC+MLM+VTM +MFM+CG
ClipBERT (Lei et al., 2021b)		2D CNN	Emb.	✗	✓	MLM+VTM
VLM (Xu et al., 2021a)		3D CNN	Emb.	✗	✗	MLM-MFM+MMM
DeCEMBERT (Tang et al., 2021b)		2D+3D CNN	Emb.	✗	✗	MLM+VTM+CA
TACo (Yang et al., 2021b)		2D+3D CNN	Xformer	✗	✗	VTM+VTC
VQA-T (Yang et al., 2021a)		3D CNN	Xformer	✗	✗	MLM+VTC
VICTOR (Lei et al., 2021a)		2D CNN	Emb.	✓	✗	MLM+VTC+MFM +FOM+SOM+CG
MERLOT (Zellers et al., 2021)		2D CNN +Xformer	Xformer	✗	✓	MLM+VTC+FOM
MV-GPT (Seo et al., 2022)		Xformer	Xformer	✓	✓	MLM+CG
HTM (Miech et al., 2019)	Dot Product	3D CNN	Word2Vec	✗	✓	VTC
MIL-NCE (Miech et al., 2020)		3D CNN	Word2Vec	✗	✓	VTC
Support Set (Patrick et al., 2020)		2D+3D CNN +Xformer	Xformer	✓	✗	VTC+CG
Frozen (Bain et al., 2021)		Xformer	Xformer	✗	✓	VTC
VideoCLIP (Xu et al., 2021b)		3D CNN +Xformer	Xformer	✗	✗	VTC

Video-Text Contrastive Learning (VTC)

- Borrow the idea from contrastive learning
- VTC aims to learn the correspondence between video and text

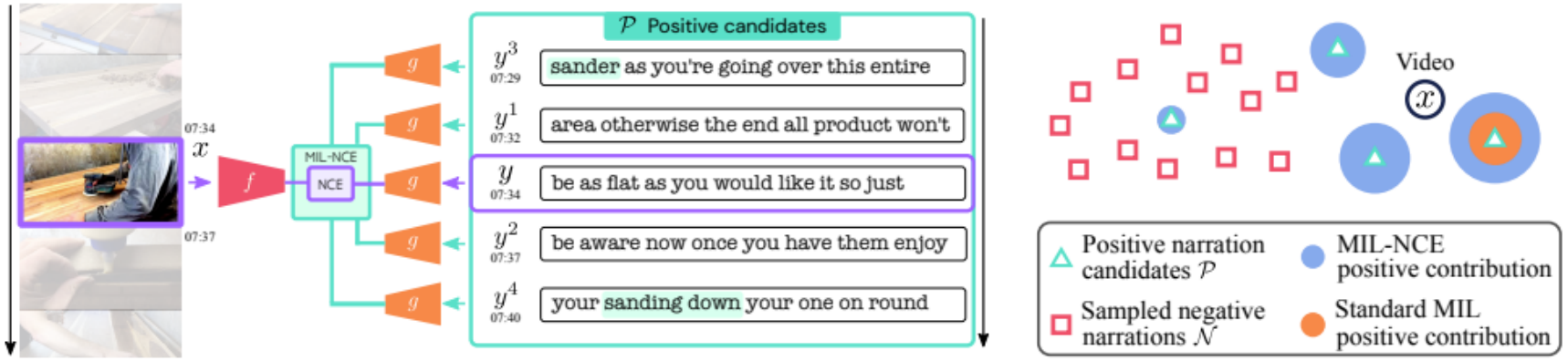


Many follow-up works propose to collect better positive and negative pairs

Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, ICCV 2019

MIL-NCE

- Multiple Instance Learning (MIL) and Noise Contrastive Estimation (NCE)
- Try to mitigate the misalignment between video and transcript
- *Consider a set of multiple positive candidate pairs*



(a) **Examples of positive candidates** (b) **Illustration of *MIL-NCE***

Figure 2: **Left.** Our MIL-NCE makes it possible to consider a set of multiple positive candidate pairs $\{(x, y), (x, y^1), \ldots, (x, y^4)\}$ while the standard NCE approach would only consider the single (x, y) training pair and miss the visually grounded object description `sander` from pair (x, y^3) or the action description `sanding down` from (x, y^4). **Right.** Given a video x and an associated set of positive narration candidates $\mathcal{P}$ (green triangles) that may or may not be correct, our *MIL-NCE* selects *multiple* correct positives (large blue areas) while downweighting incorrect positives (smaller blue areas) based on a discriminative ratio against negatives $\mathcal{N}$ (red squares). In contrast, traditional MIL considers only one positive (orange circle) while discarding the rest.

End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

Masked Language Modeling (MLM)

- MLM is a direct adoption from NLP field
- Facilitate the multimodal fusion between video and text

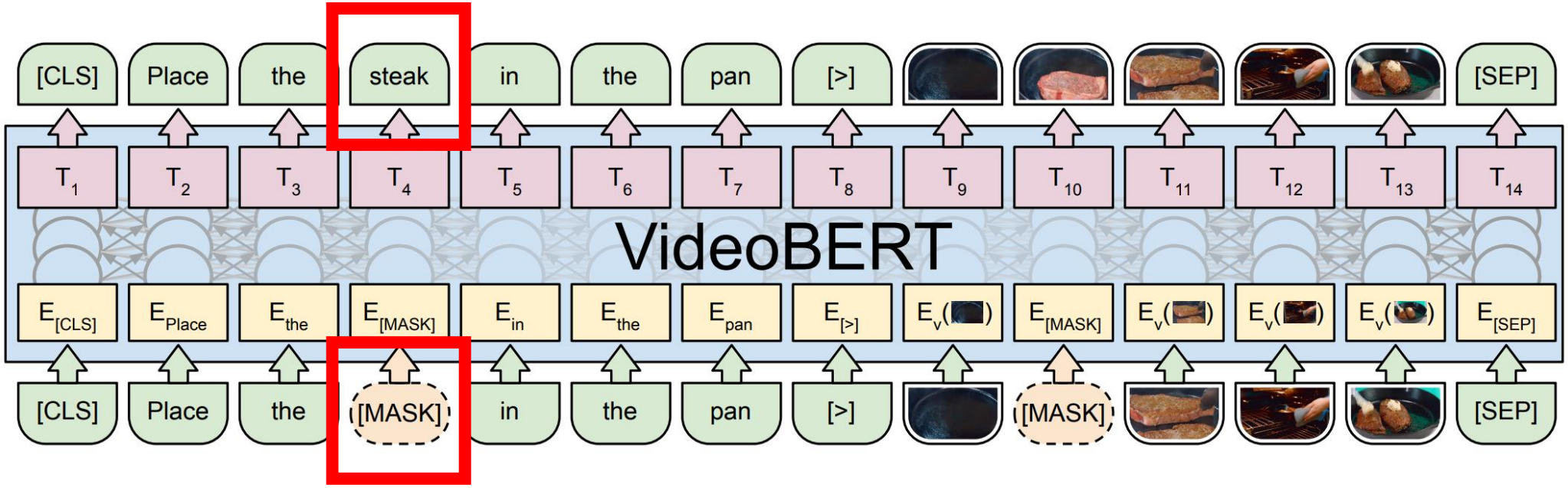


Figure credit: VideoBERT: A Joint Model for Video and Language Representation Learning, ICCV 2019

Video-Text Matching (VTM)

- Given a batch of positive and negative video-text pairs, VTM aims to identify which videos and texts correspond to each other.
- Often formulate as a binary classification task

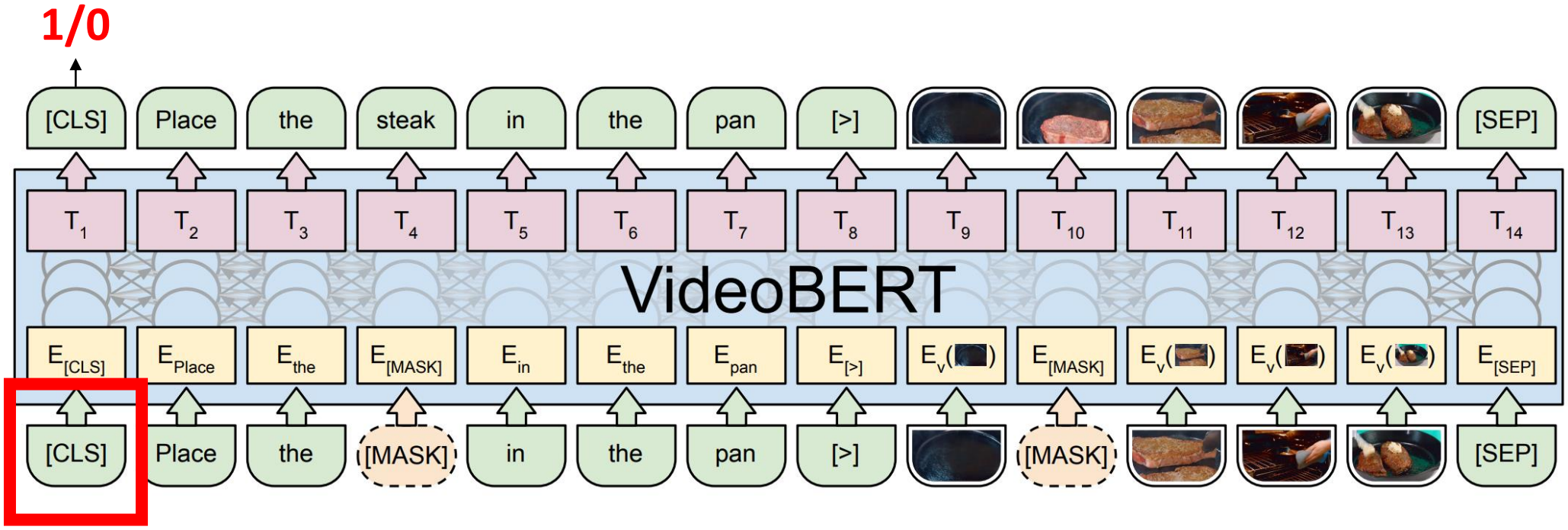
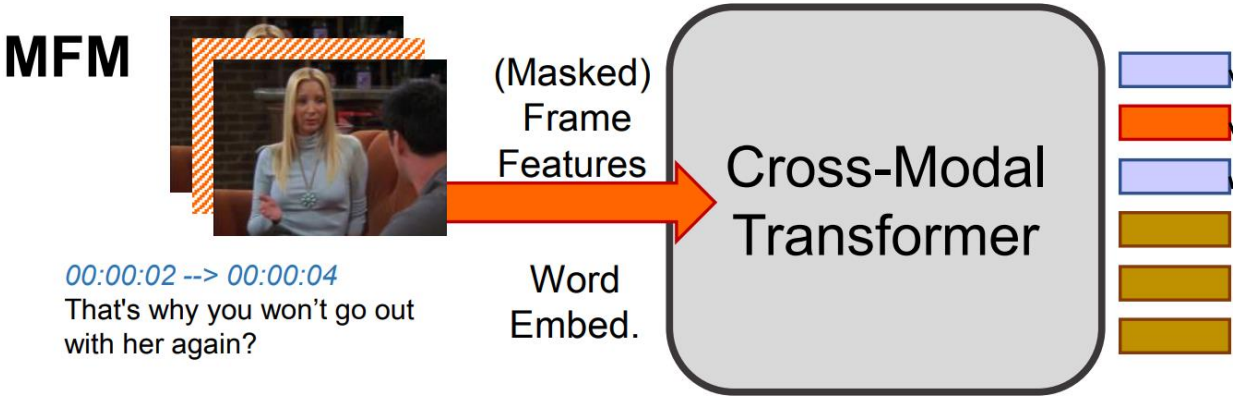


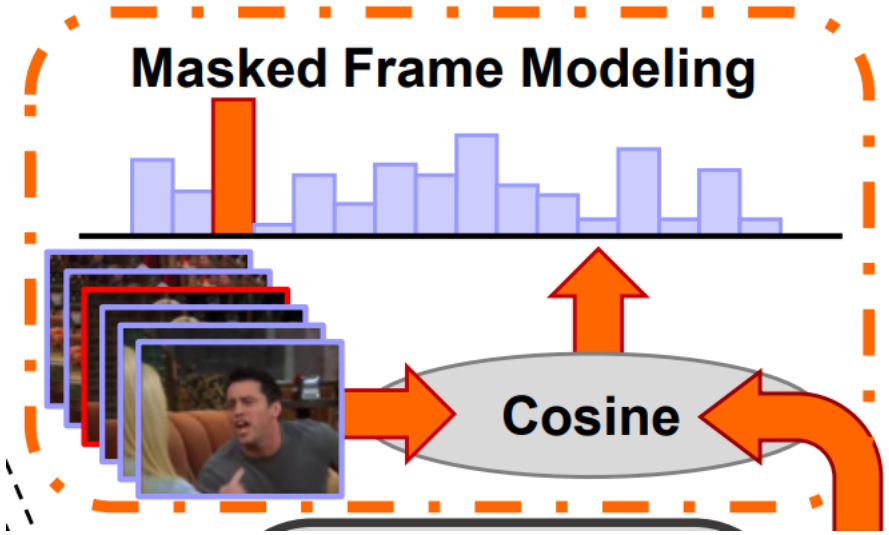
Figure credit: VideoBERT: A Joint Model for Video and Language Representation Learning, ICCV 2019

Masked Video Modeling (MVM)

- Similar to MLM, MVM is also developed to reconstruct the masked input visual tokens
- Visual features are high-dimensional and continuous
- Little-to-none effects in the pre-training



Visual Feature Regression



Noise Contrastive Estimation loss

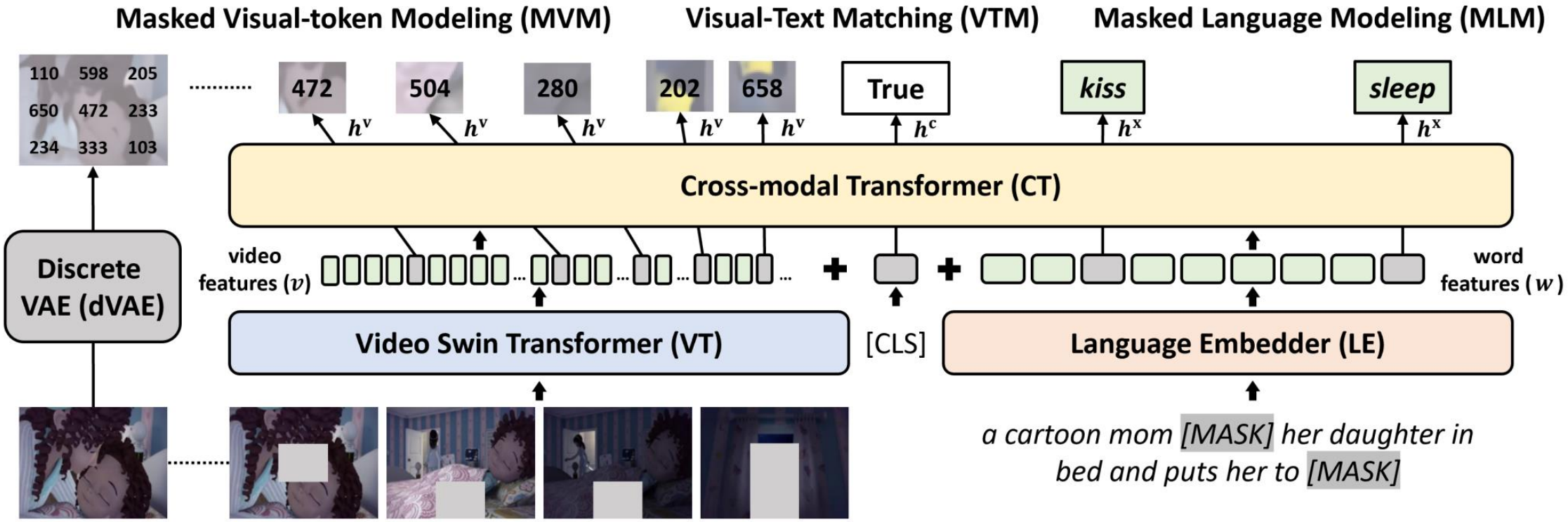
Hero: Hierarchical encoder for video+language omni-representation pre-training, EMNLP 2020
ViLT: Vision-and-language transformer without convolution or region supervision, ICML 2021

Outline

- Data and challenges
- Pioneer work in video-text pre-training
- Advanced pre-training tasks
- Transferring image-text model

Masked Visual-token Modeling

- Reconstruct the discrete latent codes from pre-trained DALL-E
- Promising improvements for video-and-language pre-training



VIOLET : End-to-End Video-Language Transformers with Masked Visual-token Modeling, arXiv 2021

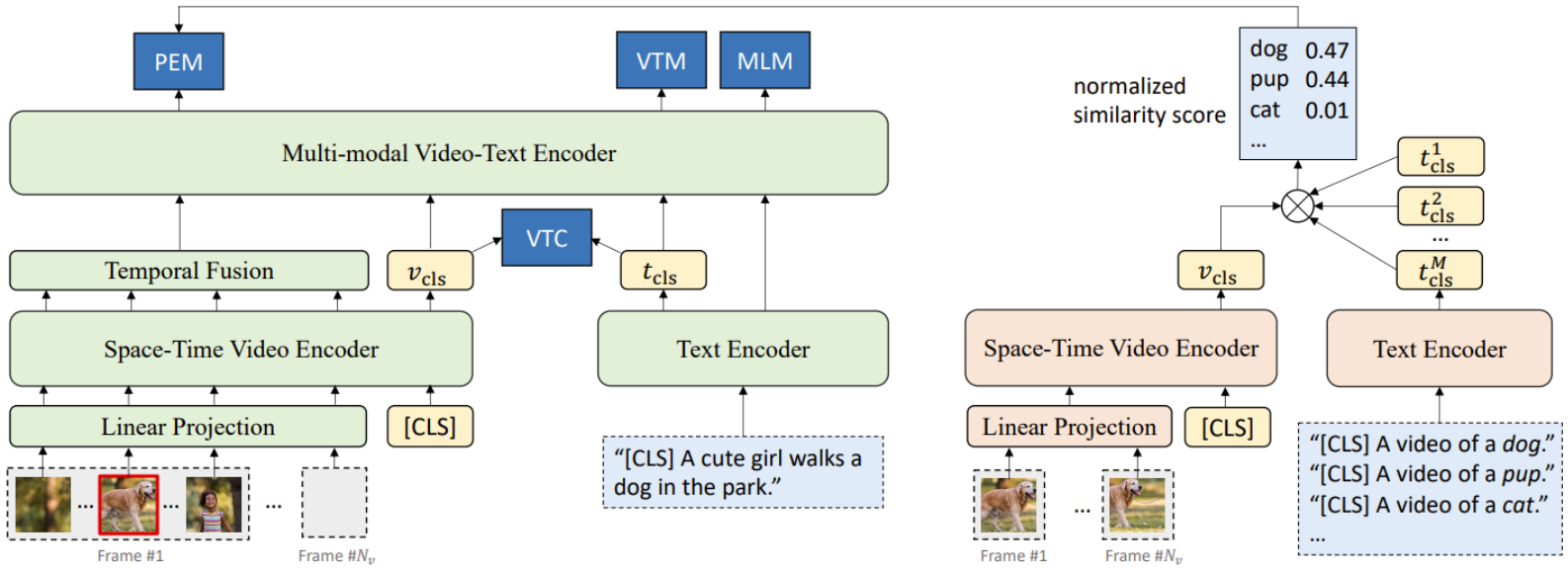
Frame Order Modeling (FOM)

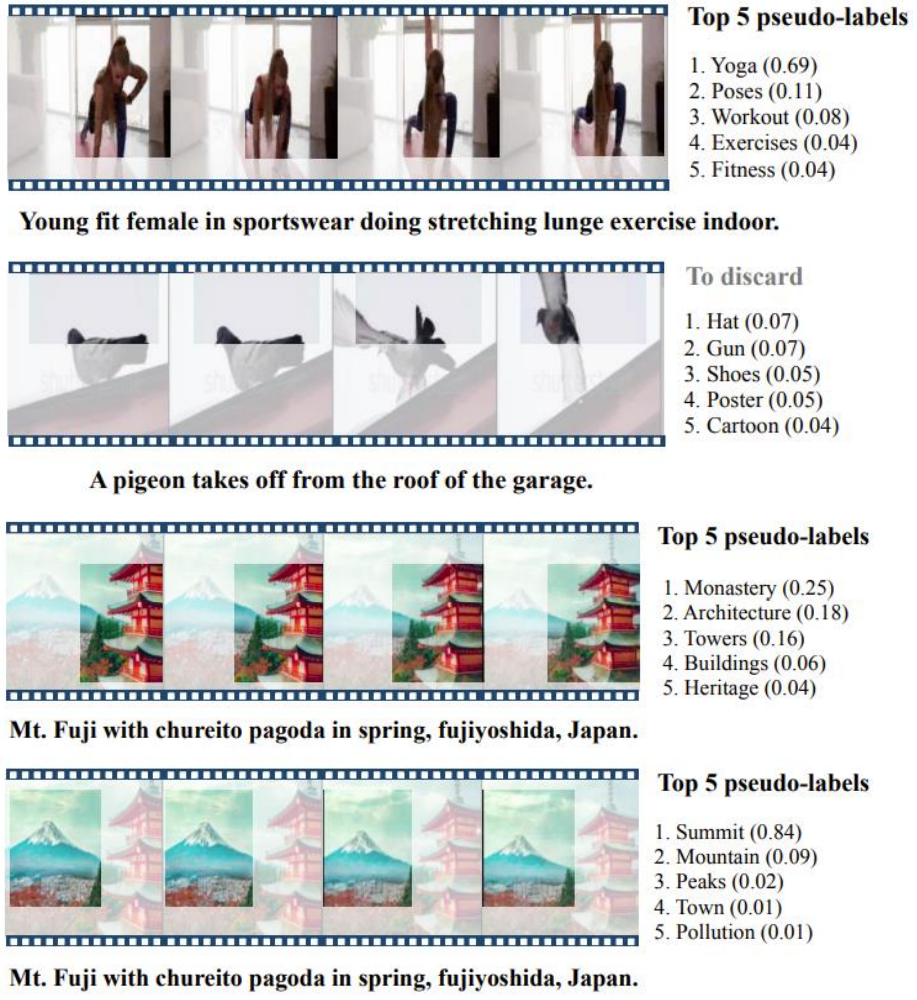
- During training, a percentage of the frames is randomly selected to be shuffled, and the goal is to reconstruct their original temporal order
- Formulate FOM as a classification task and predict the timestamp.

Hero: Hierarchical encoder for video+language omni-representation pre-training, EMNLP 2020
Merlot: Multimodal neural script knowledge models, NeurIPS 2021

Object-level Supervision

- Object-level supervision can enhance cross-modality alignment





It is helpful to learn fine-grained region-entity alignment

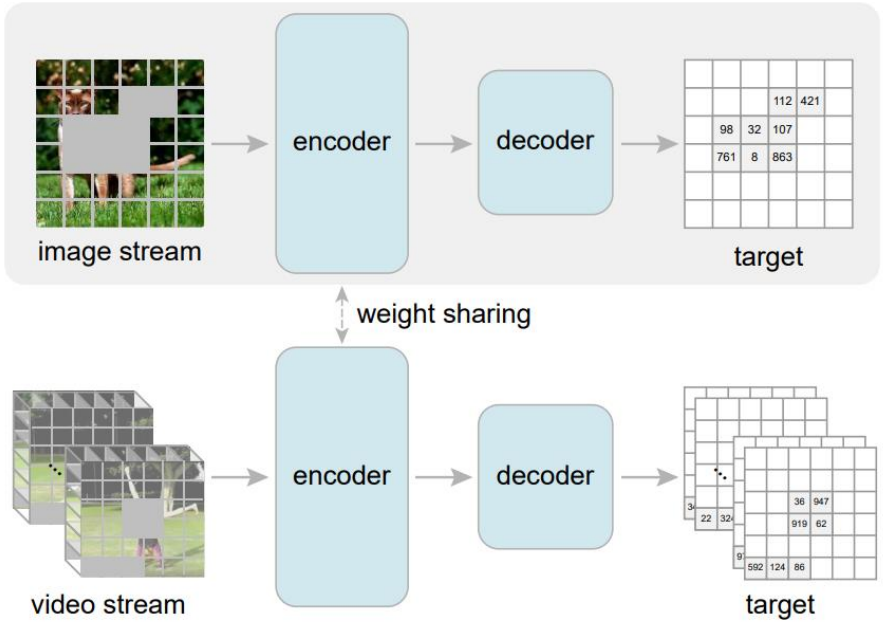
Align and Prompt: Video-and-Language Pre-training with Entity Prompts, CVPR 2022
Object-aware Video-language Pre-training for Retrieval, CVPR 2022
Actbert: Learning global-local video-text representations, CVPR 2020

Outline

- Data and challenges
- Pioneer work in video-text pre-training
- Advanced pre-training tasks
- Transferring image-text model

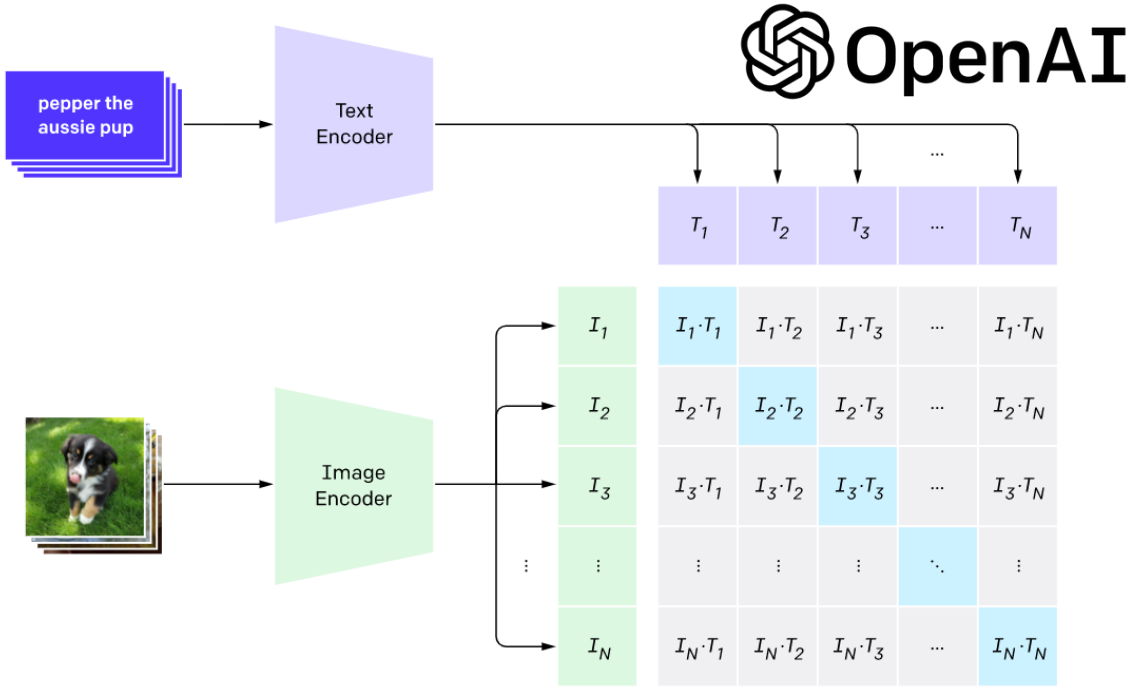
Transferring Image-Text Model

- In core video problems, leveraging ImageNet pre-trained weights as an initialization is usually helpful



BEVT: BERT Pretraining of Video Transformers, CVPR 2022
Video Swin Transformer, ICCV 2021
VidTr: Video Transformer Without Convolutions, ICCV 2021
Mask2Former for Video Instance Segmentation, ArXiv 2021

Can we leverage well pre-trained image-text model for video-text tasks?



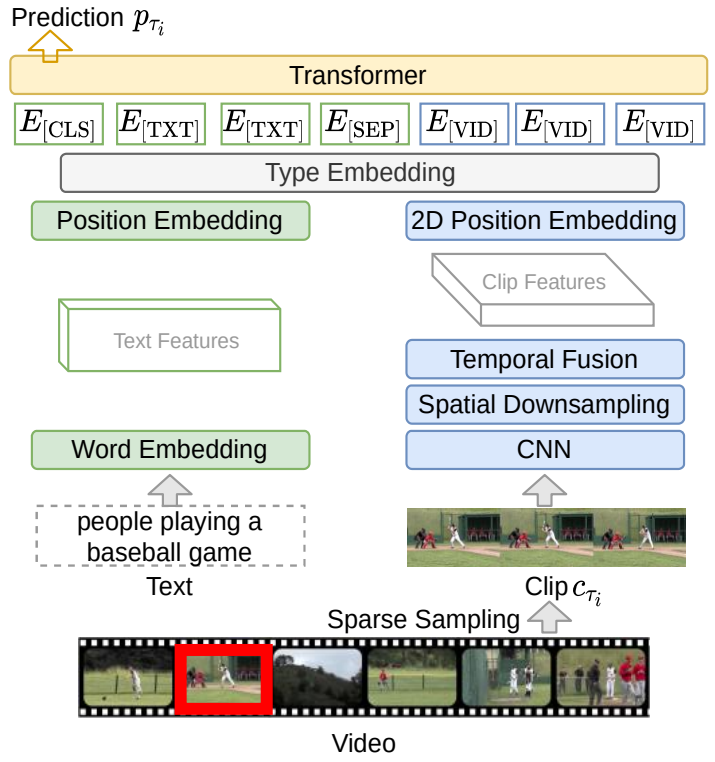
ClipBERT

- Pre-train with MLM + ITM on image-text pairs (COCO + VG Captions)
- Avoid the excessive cost of video-text pre-training

Weight Initialization		MSRVTT Retrieval				MSRVTT-QA Acc.
CNN	transformer	R1	R5	R10	MdR	
random	random	0.3	0.4	0.9	506.0	28.05
random	$\text{BERT}_{\text{BASE}}$	0.0	0.2	0.7	505.0	31.72
TSN, K700	$\text{BERT}_{\text{BASE}}$	5.7	22.1	33.1	23.0	35.40
ImageNet	$\text{BERT}_{\text{BASE}}$	7.2	23.3	35.6	21.0	35.01
grid-feat	$\text{BERT}_{\text{BASE}}$	7.4	21.0	30.7	26.0	35.27
image-text pre-training		**10.2**	**28.6**	**40.5**	**17.0**	**35.73**

Table 5: Impact of **weight initialization strategy**.

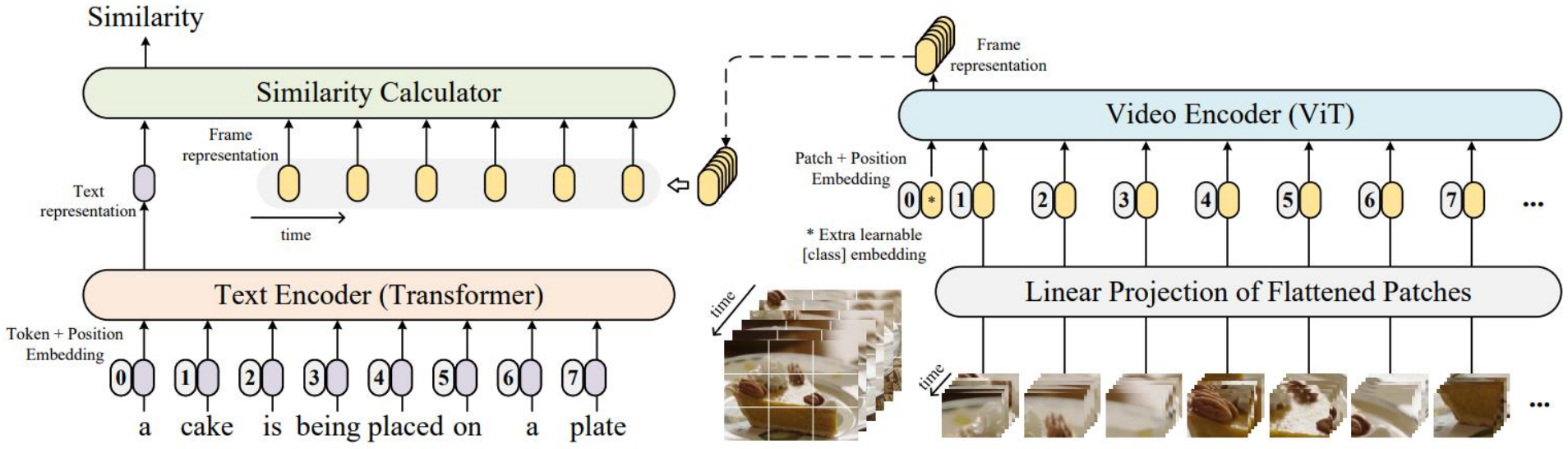
Image-text pre-training helps video-text tasks!



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

CLIP for X

- CLIP4Clip is post-pretrained with contrastive loss on HT100M



Large-scale image-text pre-training also helps video-text tasks

CLIP4Caption: CLIP for Video Caption , arXiv 2021
CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval, arXiv 2021

TubeDETR

- DETR style architecture for spatio-temporal video grounding
- Image-text pre-training (COCO, VG, F30K)

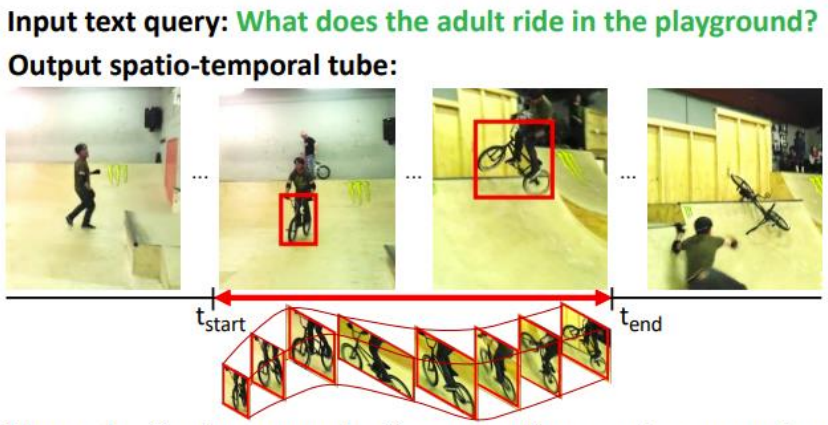


Figure 1. Spatio-temporal video grounding requires reasoning about space, time, and language.

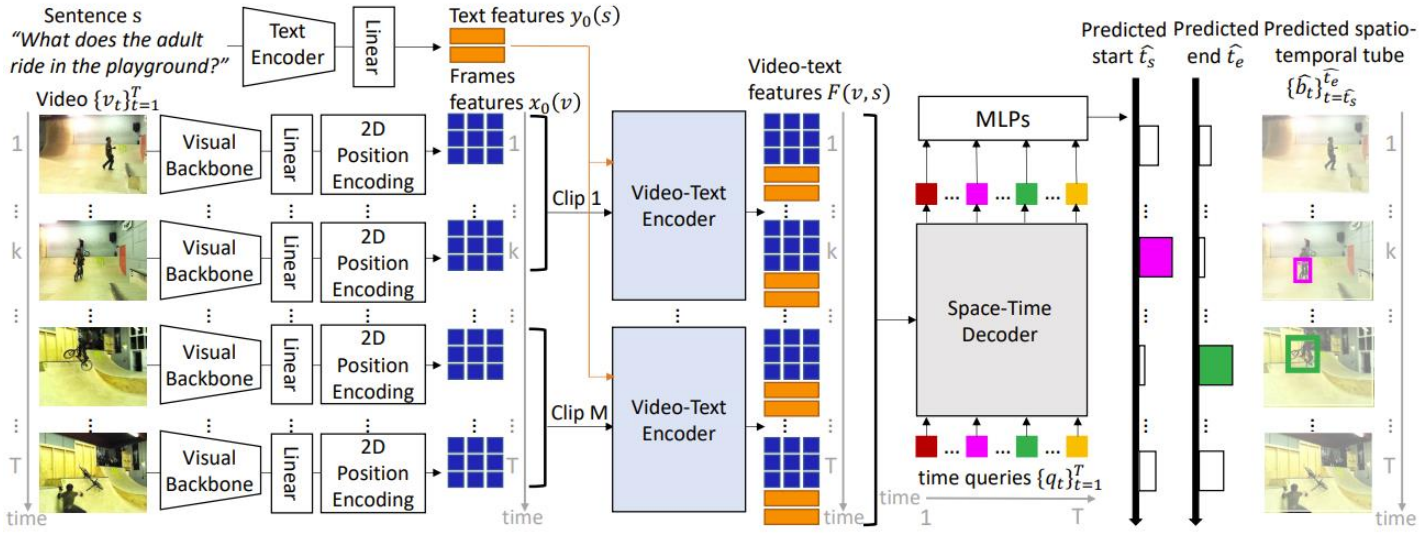


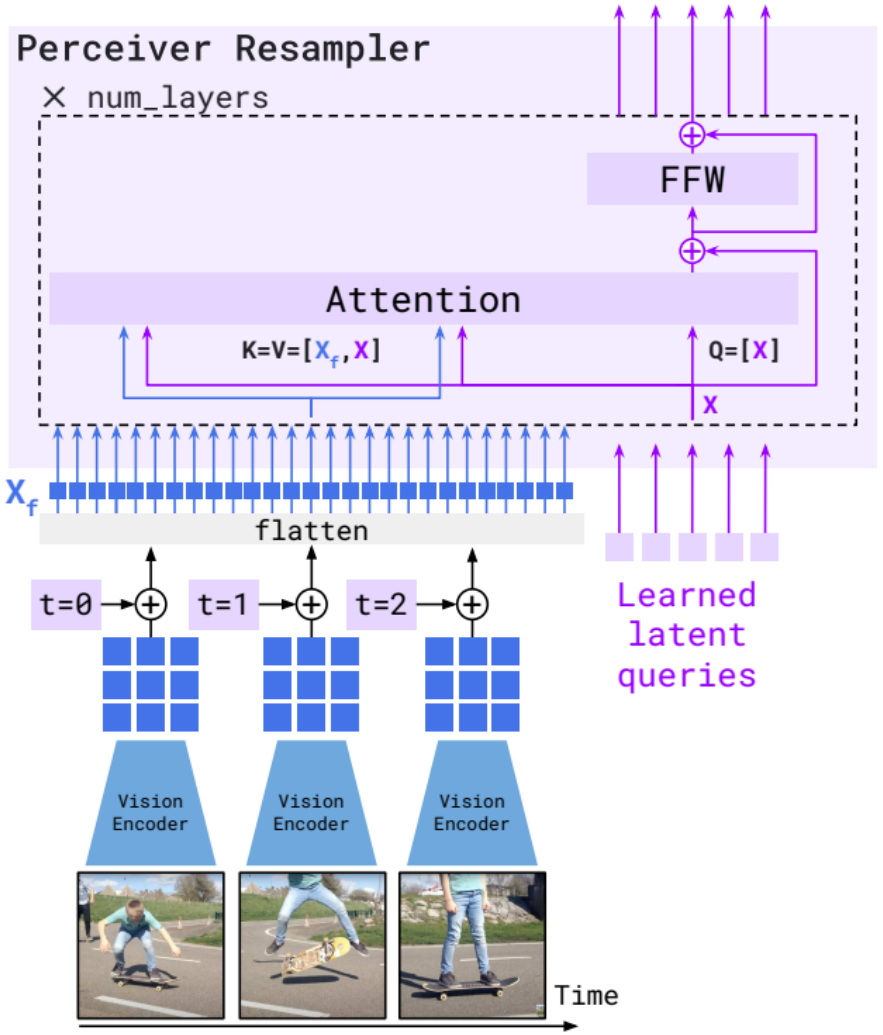
Image-text pre-training can also help advanced video-text downstream tasks

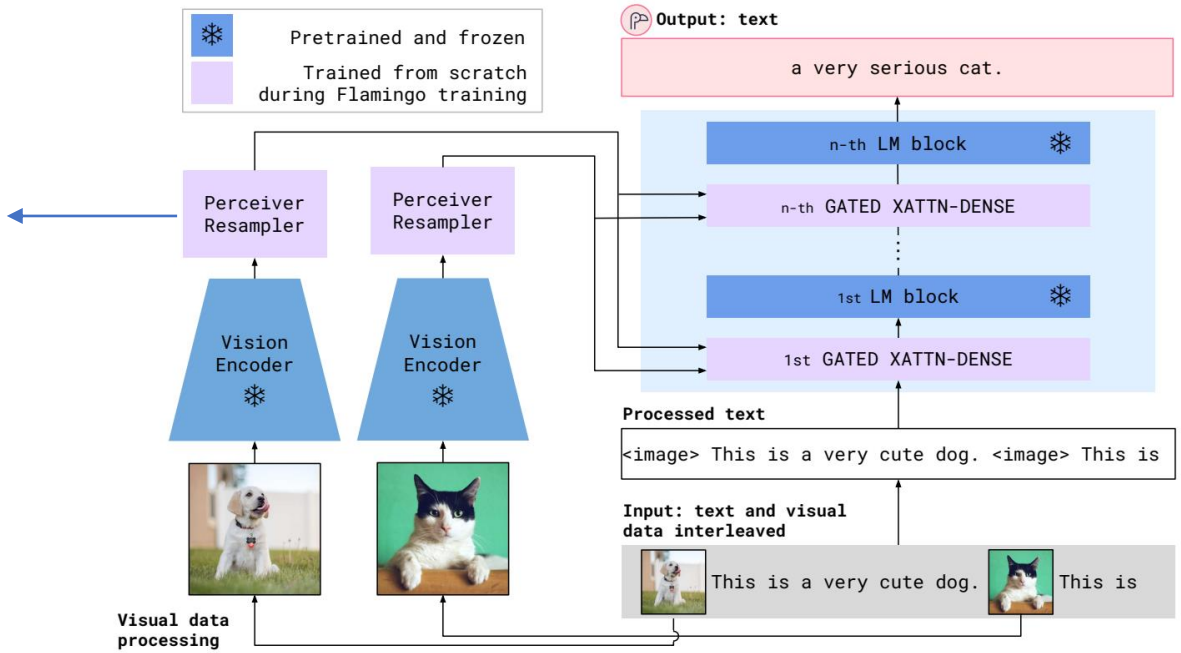
	Pre-Training	Decoder Self-Attention Transfer	m_tIoU	m_vIoU	vIoU @0.3	vIoU @0.5	m_sIoU
1.	✗	✗	42.8	23.5	33.2	20.9	38.5
2.	✓	✗	43.8	28.6	39.8	27.3	46.6
3.	✓	Temporal	**45.9**	**30.3**	**42.3**	**29.8**	**47.7**

Table 2. Effect of the weight initialization for our model on the VidSTG validation set.

TubeDETR: Spatio-Temporal Video Grounding with Transformers, CVPR 2022

Flamingo with Perceiver Resampler





A flexible architecture that can take both images and videos as inputs

Flamingo: a Visual Language Model for Few-Shot Learning, ArXiv 2022

Applying GIT to Video Domain

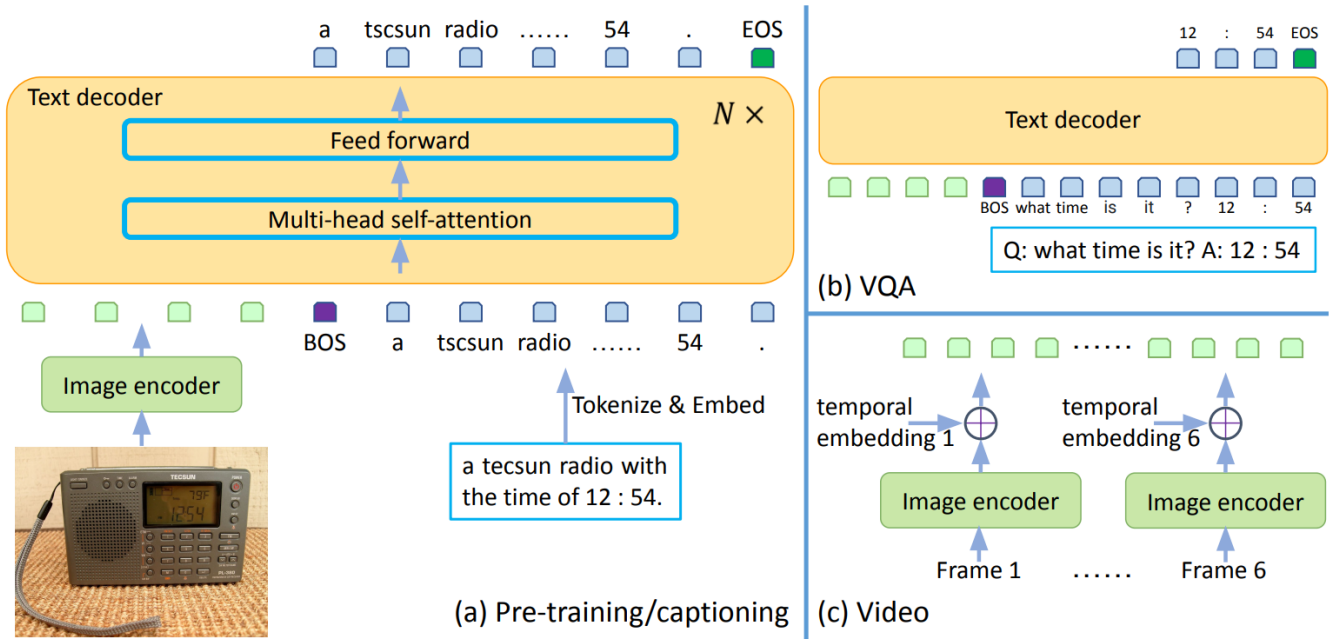
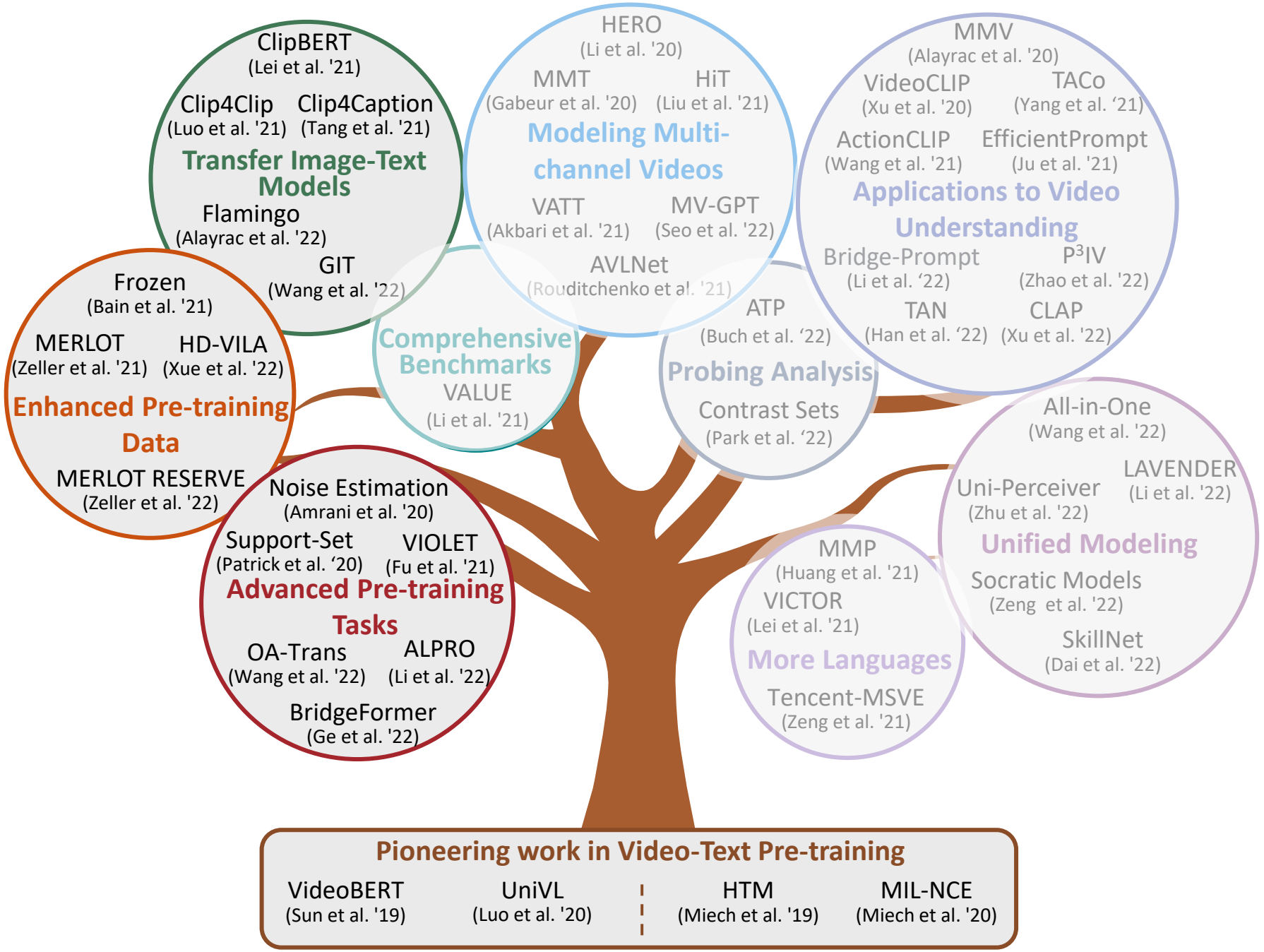


Figure 2: Network architecture of our GIT, composed of one image encoder and one text decoder. (a): The training task in both pre-training and captioning is the language modeling task to predict the associated description. (b): In VQA, the question is placed as the text prefix. (c): For video, multiple frames are sampled and encoded independently. The features are added with an extra learnable temporal embedding (initialized as 0) before concatenation.

	Video captioning			Video QA	
	MSVD	MSRVTT	VATEX*	MSVD-QA	TGIF-Frame
Prior SOTA	120.6 [60]	60 [80]	86.5 [88]	48.3 [91]	69.5 [112]
GIT (ours)	180.2	73.9	93.8	56.8	72.8
Δ	+59.6	+13.9	+7.3	+8.5	+3.3

Adaptation with sparsely-sampled frames can generate new SOTA on popular benchmark

GIT: A Generative Image-to-text Transformer for Vision and Language, ArXiv 2022


ClipBERT
(Lei et al. '21)
Clip4Clip
(Luo et al. '21)
Clip4Caption
(Tang et al. '21)
Transfer Image-Text Models
Flamingo
(Alayrac et al. '22)
GIT
(Wang et al. '22)
HERO
(Li et al. '20)
MMT
(Gabeur et al. '20)
HiT
(Liu et al. '21)
Modeling Multi-channel Videos
VATT
(Akbari et al. '21)
MV-GPT
(Seo et al. '22)
AVLNet
(Rouditchenko et al. '21)
MMV
(Alayrac et al. '20)
VideoCLIP
(Xu et al. '20)
TACo
(Yang et al. '21)
ActionCLIP
(Wang et al. '21)
EfficientPrompt
(Ju et al. '21)
Applications to Video Understanding
Bridge-Prompt
(Li et al. '22)
P3IV
(Zhao et al. '22)
TAN
(Han et al. '22)
CLAP
(Xu et al. '22)
Frozen
(Bain et al. '21)
MERLOT
(Zeller et al. '21)
HD-VILA
(Xue et al. '22)
Enhanced Pre-training Data
MERLOT RESERVE
(Zeller et al. '22)
Comprehensive Benchmarks
VALUE
(Li et al. '21)
ATP
(Buch et al. '22)
Probing Analysis
Contrast Sets
(Park et al. '22)
All-in-One
(Wang et al. '22)
Uni-Perceiver
(Zhu et al. '22)
LAVENDER
(Li et al. '22)
Unified Modeling
Socratic Models
(Zeng et al. '22)
SkillNet
(Dai et al. '22)
Noise Estimation
(Amrani et al. '20)
Support-Set
(Patrick et al. '20)
VIOLET
(Fu et al. '21)
Advanced Pre-training Tasks
OA-Trans
(Wang et al. '22)
ALPRO
(Li et al. '22)
BridgeFormer
(Ge et al. '22)
MMP
(Huang et al. '21)
VICTOR
(Lei et al. '21)
More Languages
Tencent-MSVE
(Zeng et al. '21)
Pioneering work in Video-Text Pre-training
VideoBERT
(Sun et al. '19)
UniVL
(Luo et al. '20)
HTM
(Miech et al. '19)
MIL-NCE
(Miech et al. '20)

Looking forward

- How to effectively transfer image-text model to video-text tasks?
  - Many recent methods only use a naïve frame concatenation

- Temporal modeling has not been well-explored
  - Most existing studies focus mainly on spatial modeling