Large-scale Video-Language Pre-training

Mike Z. SHOU

Asst Prof, National U. of Singapore

Oct 24, 2022

https://sites.google.com/view/showlab

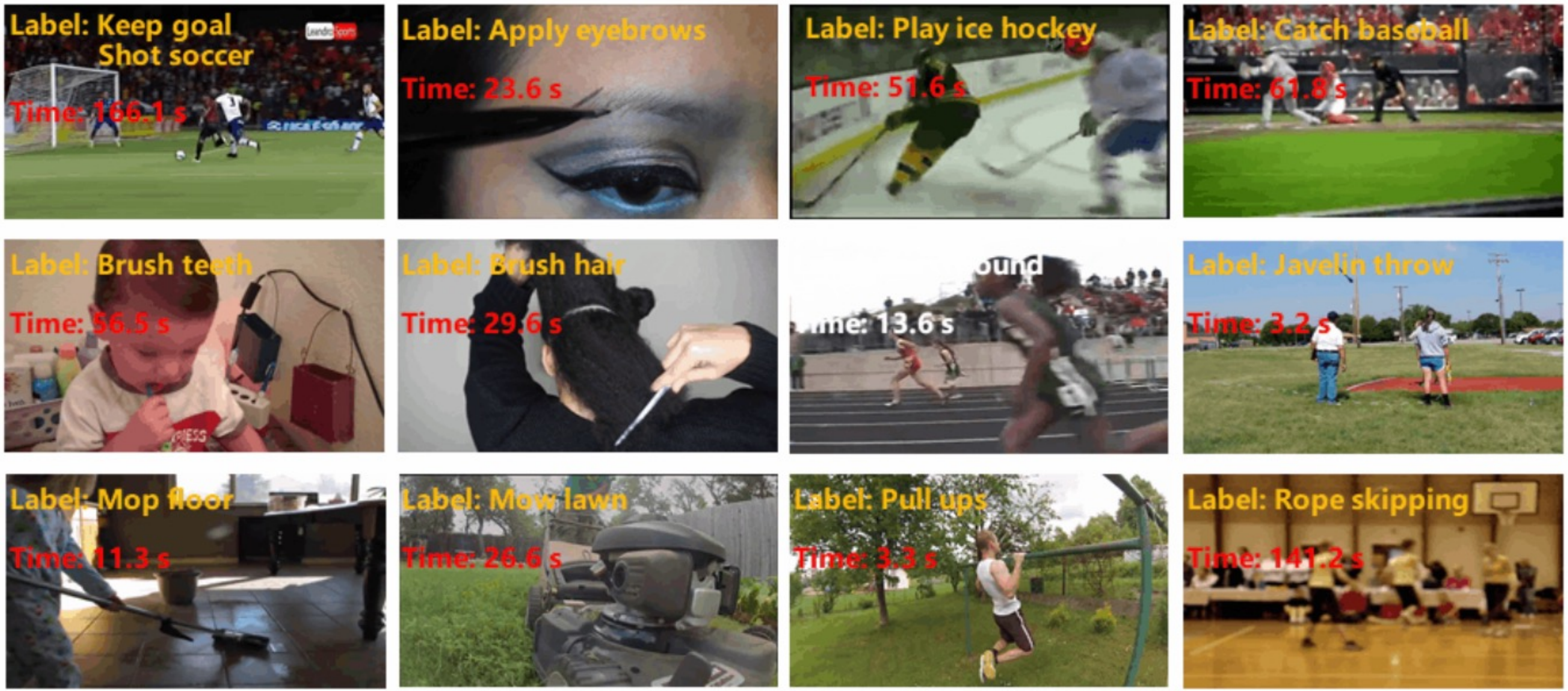
Deeper Action

NUS National University of Singapore

Show

Why large-scale pre-training?

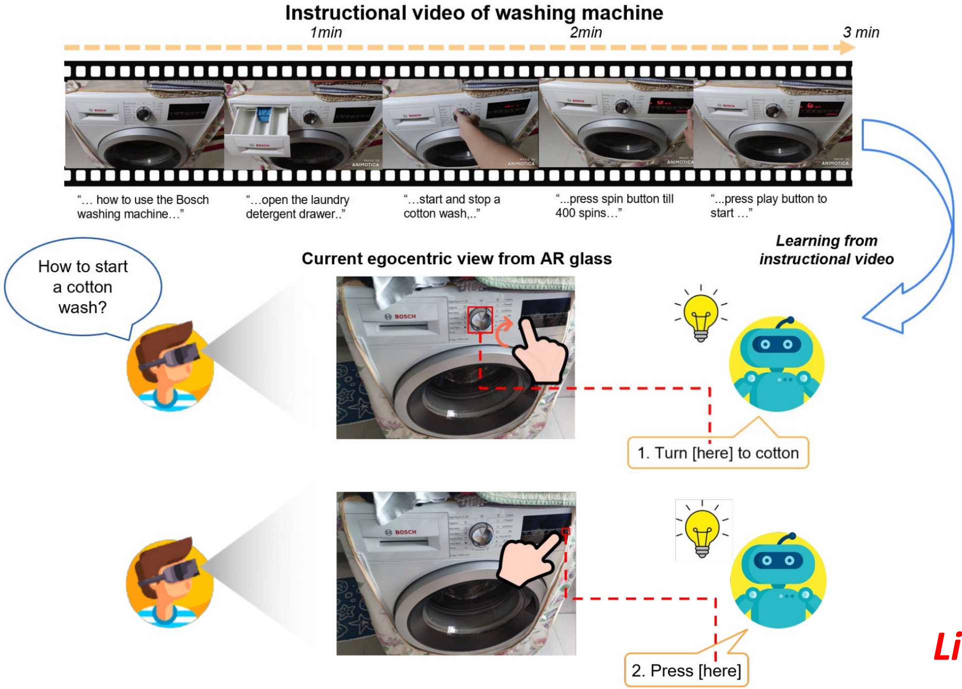
Trend: Simple action → Fine-grained action



[credit to DeeperAction Workshop]

Why large-scale pre-training?

Trend: Action classification/detection → Personal AI Assistant

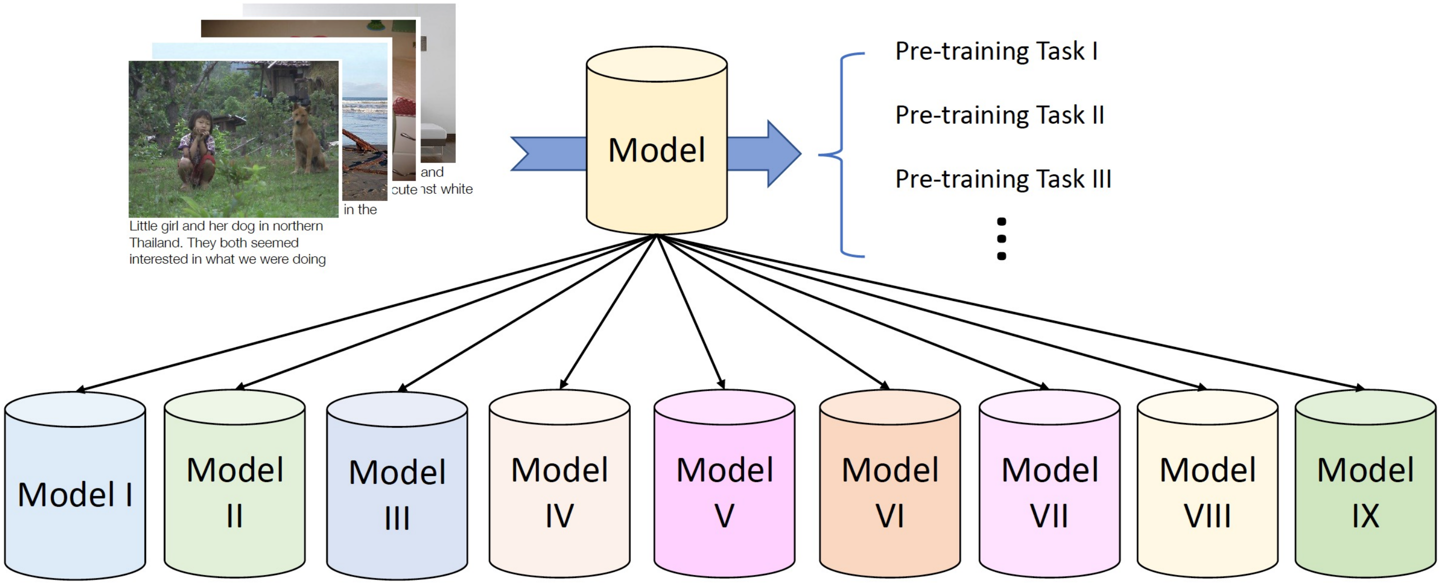

Instructional video of washing machine
1min
2min
3 min
"… how to use the Bosch washing machine…"
"…open the laundry detergent drawer.."
"…start and stop a cotton wash,.."
"...press spin button till 400 spins…"
"...press play button to start …"
Learning from instructional video
How to start a cotton wash?
Current egocentric view from AR glass
1. Turn [here] to cotton
2. Press [here]

Limited Training Data

[ECCV'22] Wong, Chen, Wu, Lei, Mao, Gao, Shou. "AssistQ: Affordance-centric Question-driven Task Completion for Egocentric Assistant".

Why large-scale Video-Language Pre-training (VLP)?

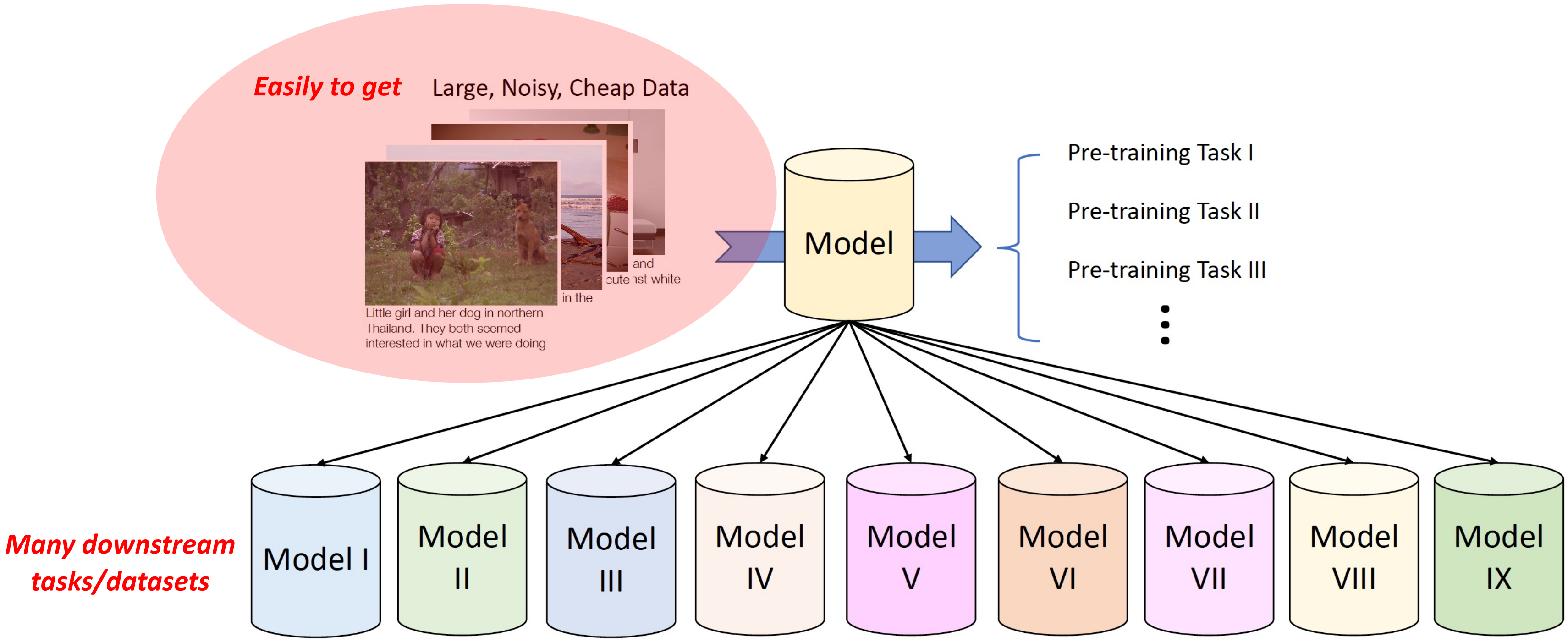
Easily to get Large, Noisy, Cheap Data



Many downstream tasks/datasets

[credit to Zhe Gan]

VLP Datasets



Easily to get Large, Noisy, Cheap Data

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing

in the

cute

and

nst white

Model

Pre-training Task I

Pre-training Task II

Pre-training Task III

⋮

Many downstream tasks/datasets

Model I

Model II

Model III

Model IV

Model V

Model VI

Model VII

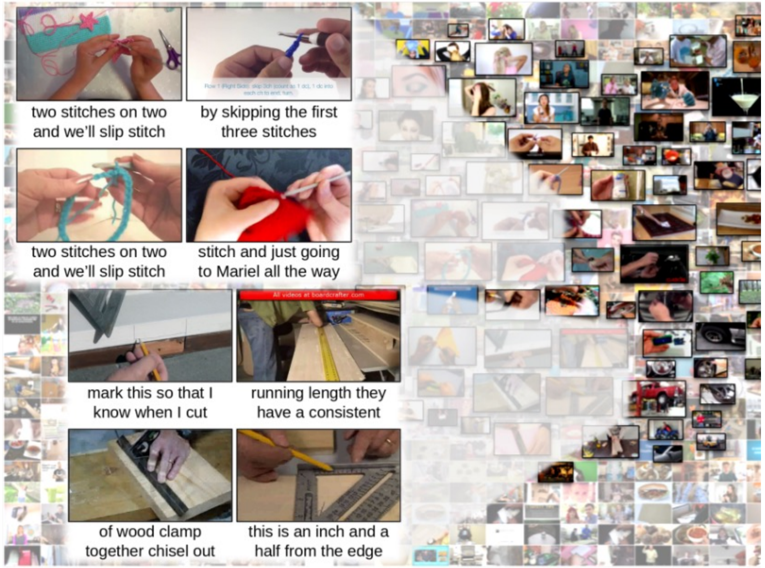
Model VIII

Model IX

[credit to Zhe Gan]

VLP Datasets

HowTo100M [ICCV 2019]
-- large, noisy


two stitches on two and we'll slip stitch
by skipping the first three stitches
two stitches on two and we'll slip stitch
stitch and just going to Mariel all the way
mark this so that I know when I cut
running length they have a consistent
of wood clamp together chisel out
this is an inch and a half from the edge

WebVid 2.5M [ICCV 2021]
-- high quality text

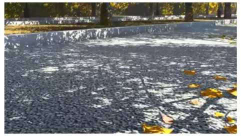
Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking.

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

Billiards, concentrated young woman playing in club.

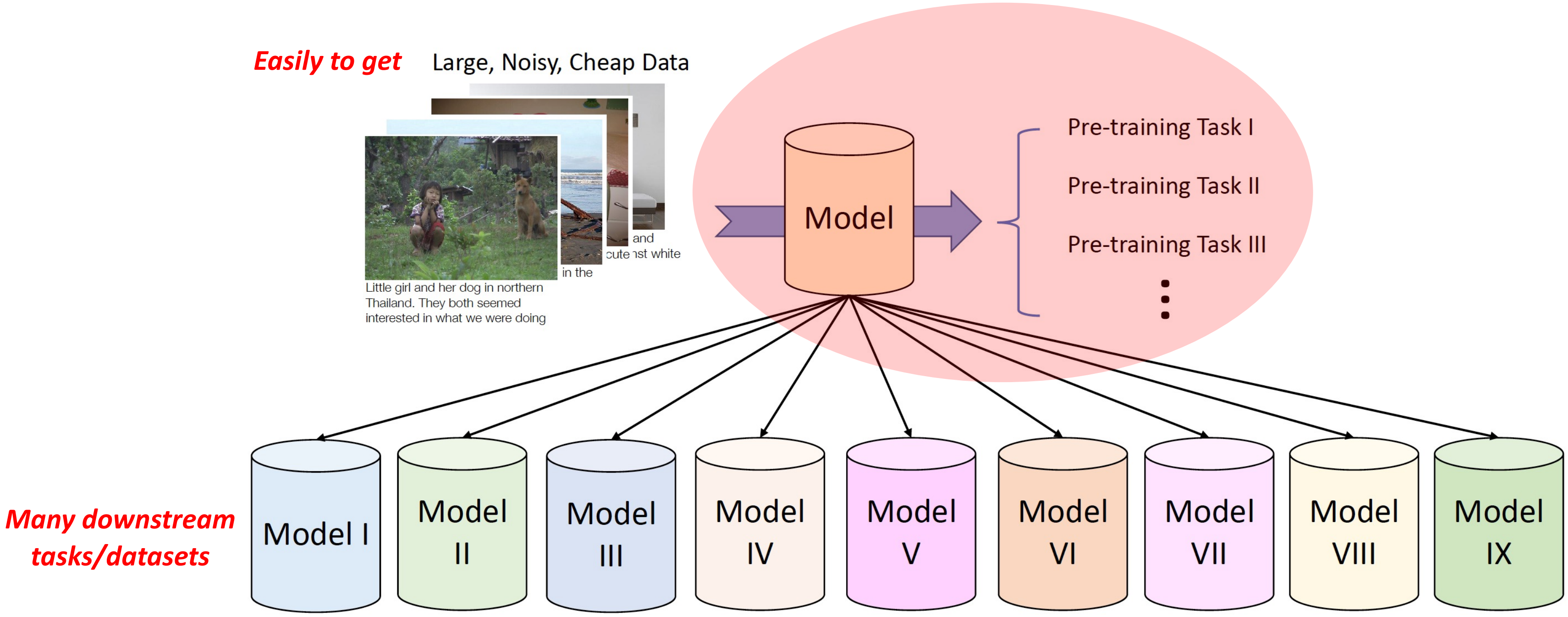
Cabeza de toro, punta cana/ dominican republic - feb 20, 2020: 4k drone flight over coral reef with manta

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



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

VLP Models



Easily to get Large, Noisy, Cheap Data

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing
in the
cute nst white
and

Model

Pre-training Task I

Pre-training Task II

Pre-training Task III

⋮

Many downstream tasks/datasets

Model I

Model II

Model III

Model IV

Model V

Model VI

Model VII

Model VIII

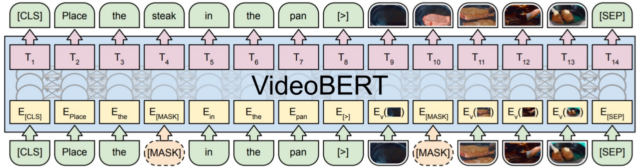
Model IX

[credit to Zhe Gan]

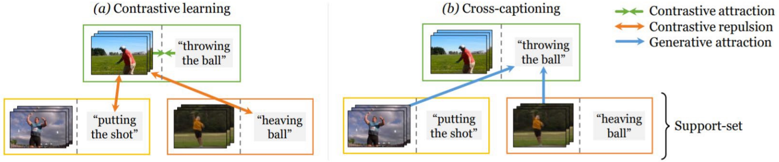
VLP Models

Early works are based on extracted features, not end-to-end

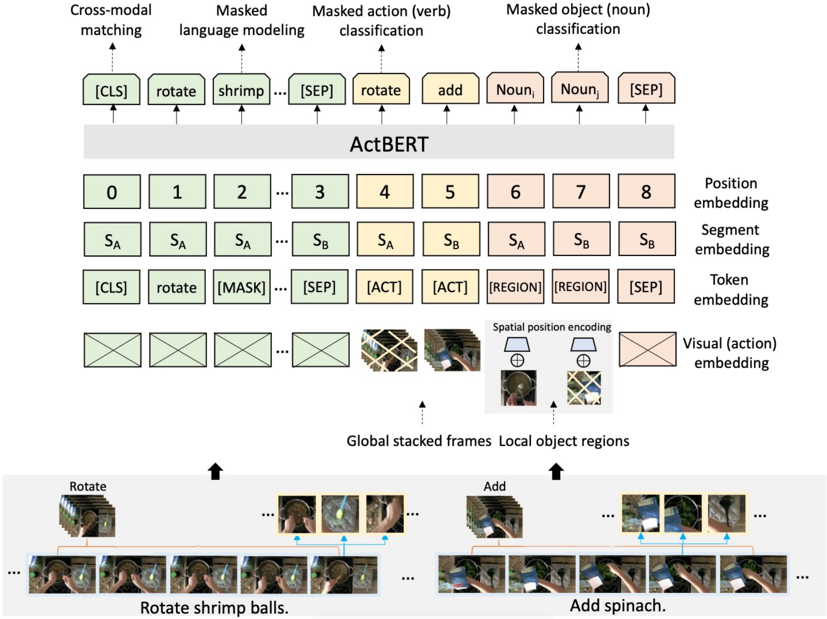
ICCV’19, Google, VideoBERT


[CLS]
Place
the
steak
in
the
pan
[>]
[SEP]
T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10} T_{11} T_{12} T_{13} T_{14}
VideoBERT
$E_{[CLS]}$ E_{Place} E_{the} $E_{[MASK]}$ E_{in} E_{the} E_{pan} $E_{[>]}$ E_v $E_{[MASK]}$ E_v E_v E_v $E_{[SEP]}$
[CLS]
Place
the
[MASK]
in
the
pan
[>]
[MASK]
[SEP]

ICLR’21, Facebook, SSB



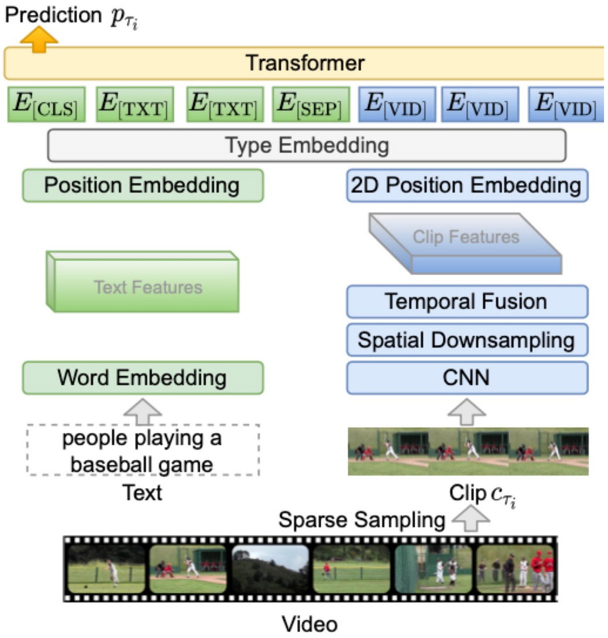
CVPR’20, UTS, ActBERT



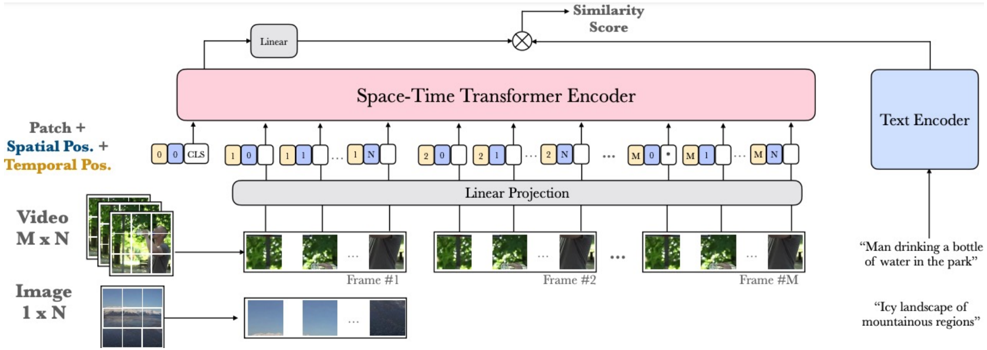
VLP Models

Better performances achieved with end-to-end training, as expected

CVPR’21, Microsoft, ClipBert



ICCV’21, VGG @ Oxford, Frozen-in-Time



Similarity Score
Linear
Space-Time Transformer Encoder
Text Encoder
Patch + Spatial Pos. + Temporal Pos.
CLS
Linear Projection
Video M x N
Image 1 x N
Frame #1
Frame #2
Frame #M
“Man drinking a bottle of water in the park”
“Icy landscape of mountainous regions”

VLP Models

Better performances achieved with end-to-end training, as expected

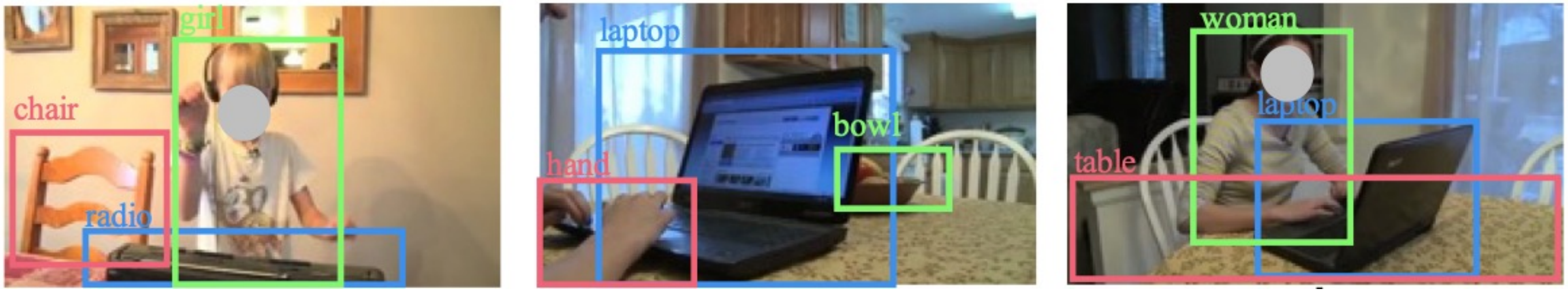
Frame-level,

No object / region info…

Modeling object in VLP?

The strong correspondence between objects in videos and in sentence

"A little girl dancing to music and a teenage girl using a computer "



Modeling object in VLP?

Modeling objects in E2E VLP -- why not video?

#1 Computational expensive:

- *10s video, even sample 1 frame per second, 10 frames*
- *For each frame, typically ~30 boxes*

#2 High redundancy over frames -- makes optimization challenging

Modeling object in VLP?

Maximize object info vs. Minimize #regions

Modeling object in VLP?

Object-aware Video-language Pre-training for Retrieval

Joint work

w/ Alex Jinpeng Wang

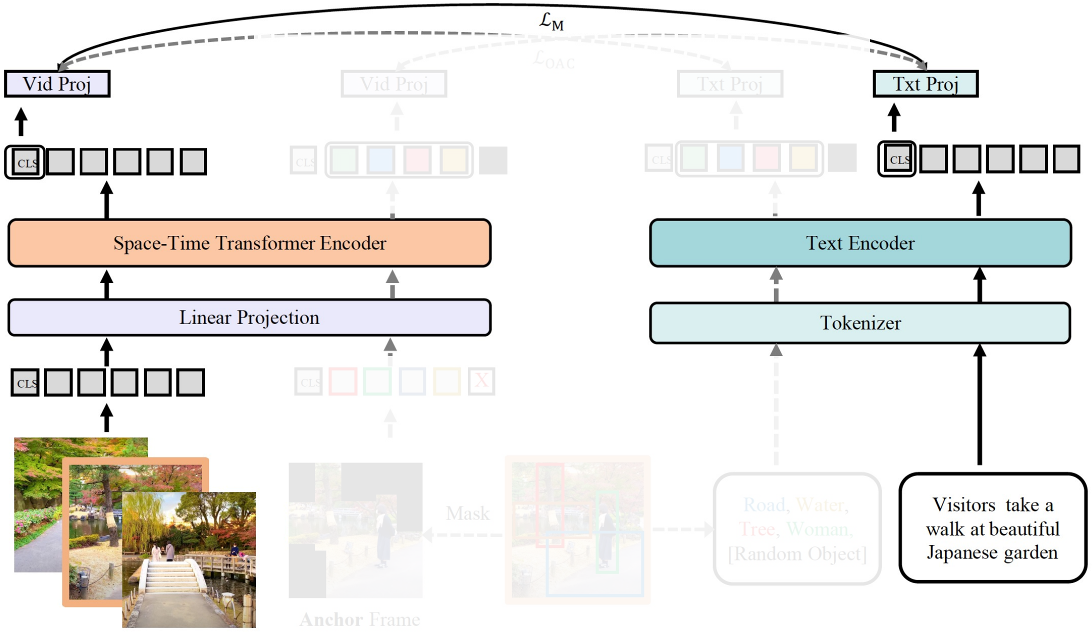
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

https://github.com/FingerRec/OA-Transformer

Object-Aware Transformer

Traditional two-stream model e2e VLP model

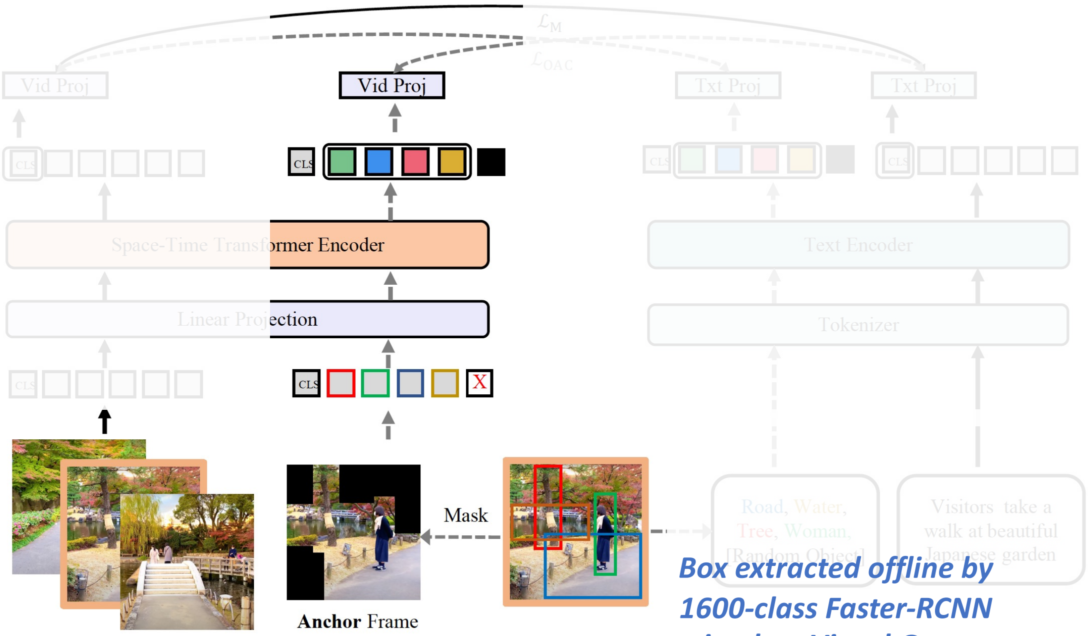
TimeSformer (12-layer ViT-B/16)



DistillBert

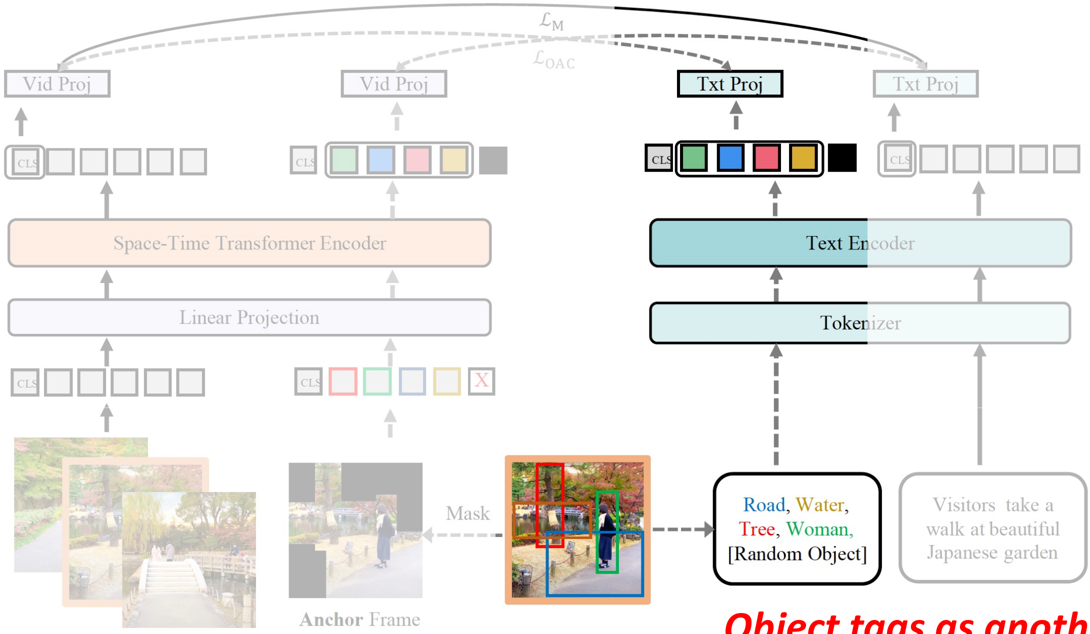
Object-Aware Transformer

1 single anchor frame for encoding object information



Box extracted offline by 1600-class Faster-RCNN trained on Visual Genome

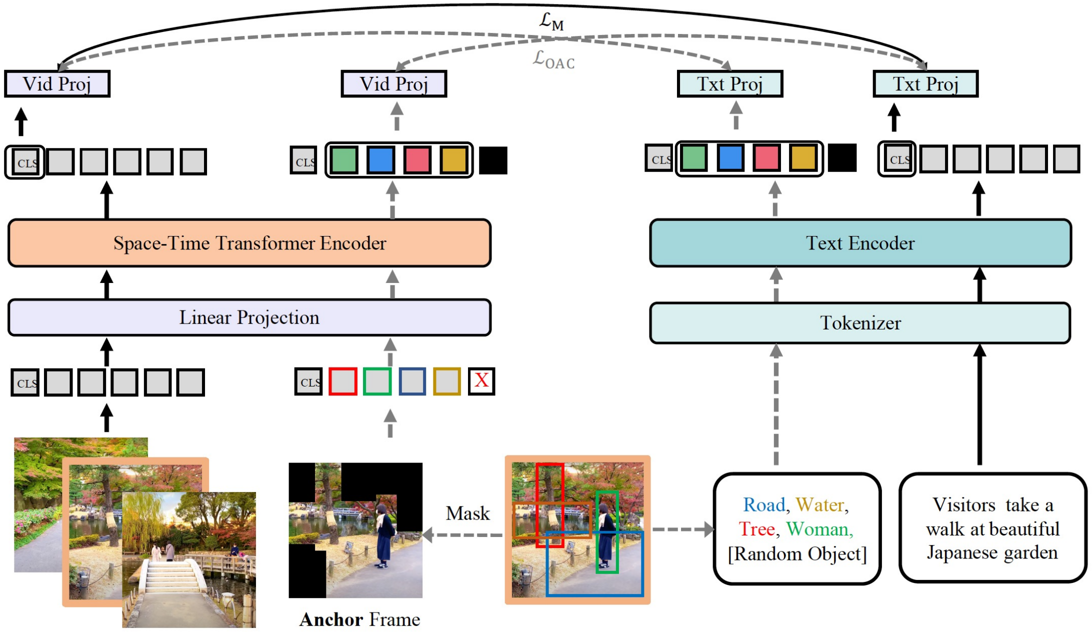
Object-Aware Transformer



Object tags as another text stream

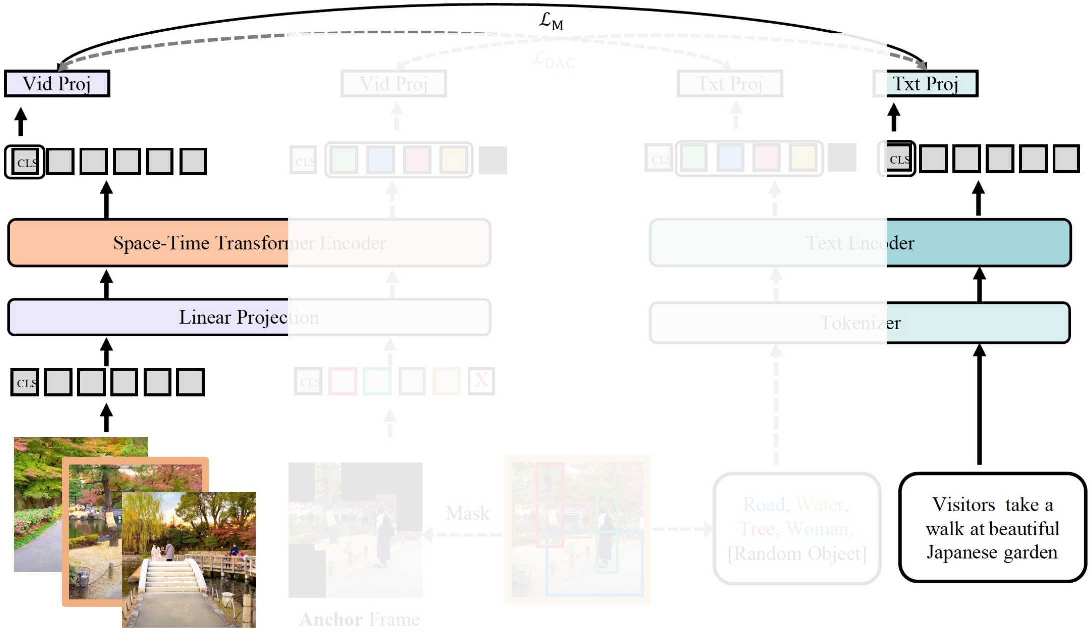
Object-Aware Transformer

Object-aware contrastive loss between 4 streams

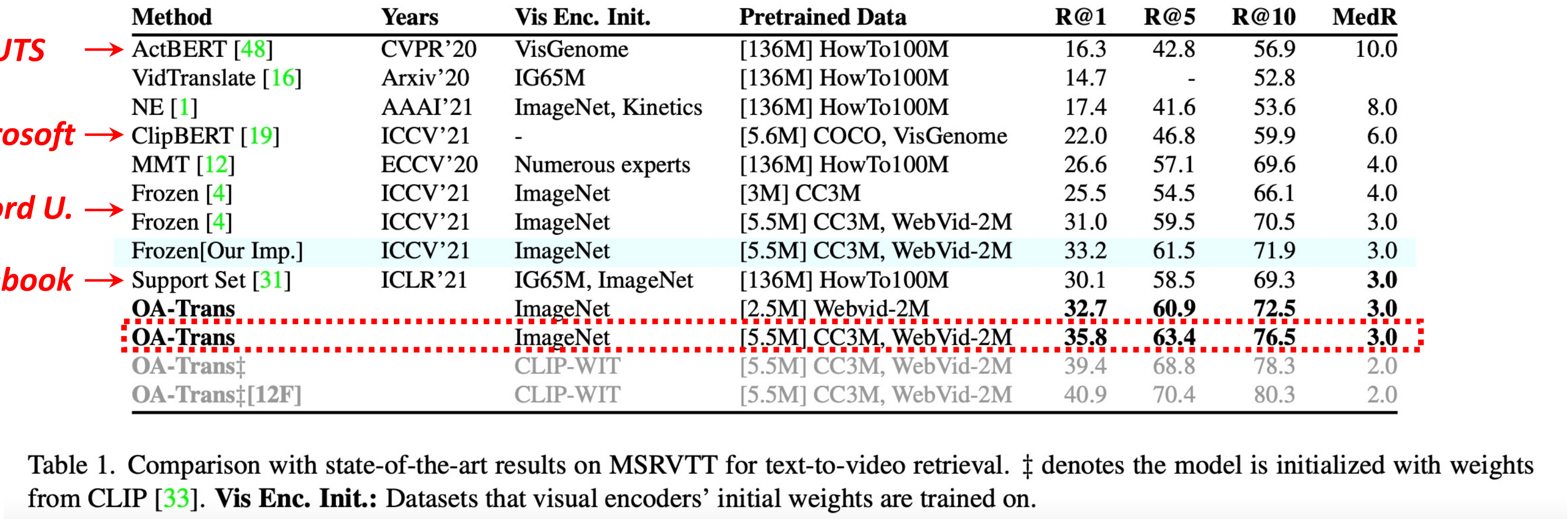


Fine-tuning & Testing

During downstream fine-tuning & inference, no need to run object detection and we remove the 2 object streams to ensure high efficiency



Comparisons with SOTA



UTS →

Microsoft →

Oxford U. →

Facebook →

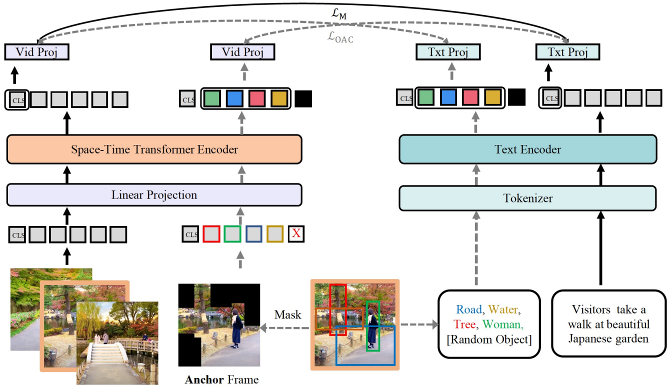
Method	Years	Vis Enc. Init.	Pretrained Data	R@1	R@5	R@10	MedR
ActBERT [48]	CVPR'20	VisGenome	[136M] HowTo100M	16.3	42.8	56.9	10.0
VidTranslate [16]	Arxiv'20	IG65M	[136M] HowTo100M	14.7	-	52.8	
NE [1]	AAAI'21	ImageNet, Kinetics	[136M] HowTo100M	17.4	41.6	53.6	8.0
ClipBERT [19]	ICCV'21	-	[5.6M] COCO, VisGenome	22.0	46.8	59.9	6.0
MMT [12]	ECCV'20	Numerous experts	[136M] HowTo100M	26.6	57.1	69.6	4.0
Frozen [4]	ICCV'21	ImageNet	[3M] CC3M	25.5	54.5	66.1	4.0
Frozen [4]	ICCV'21	ImageNet	[5.5M] CC3M, WebVid-2M	31.0	59.5	70.5	3.0
Frozen[Our Imp.]	ICCV'21	ImageNet	[5.5M] CC3M, WebVid-2M	33.2	61.5	71.9	3.0
Support Set [31]	ICLR'21	IG65M, ImageNet	[136M] HowTo100M	30.1	58.5	69.3	**3.0**
OA-Trans		ImageNet	[2.5M] Webvid-2M	**32.7**	**60.9**	**72.5**	**3.0**
OA-Trans		ImageNet	[5.5M] CC3M, WebVid-2M	**35.8**	**63.4**	**76.5**	**3.0**
OA-Trans‡		CLIP-WIT	[5.5M] CC3M, WebVid-2M	39.4	68.8	78.3	2.0
OA-Trans‡[12F]		CLIP-WIT	[5.5M] CC3M, WebVid-2M	40.9	70.4	80.3	2.0

Table 1. Comparison with state-of-the-art results on MSRVTT for text-to-video retrieval. ‡ denotes the model is initialized with weights from CLIP [33]. **Vis Enc. Init.:** Datasets that visual encoders' initial weights are trained on.

From retrieval to more tasks

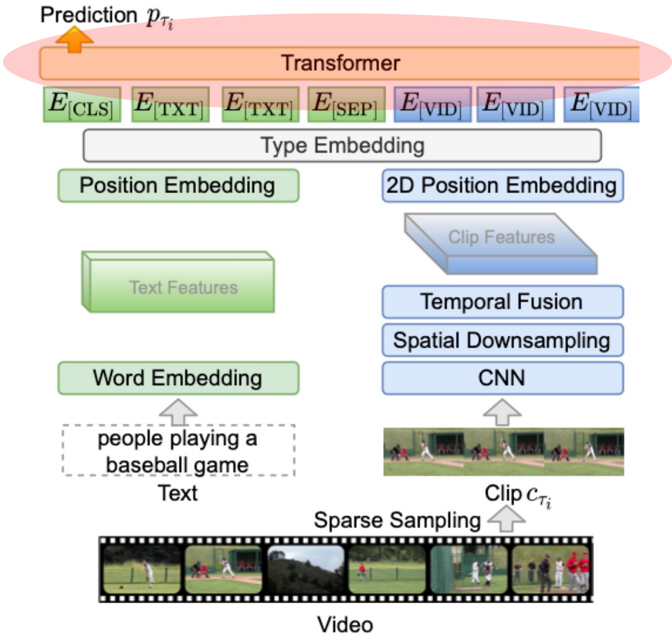
- **Good on retrieval task**
- **For other tasks like QA, need more complex fusion**

Object-Aware Transformer

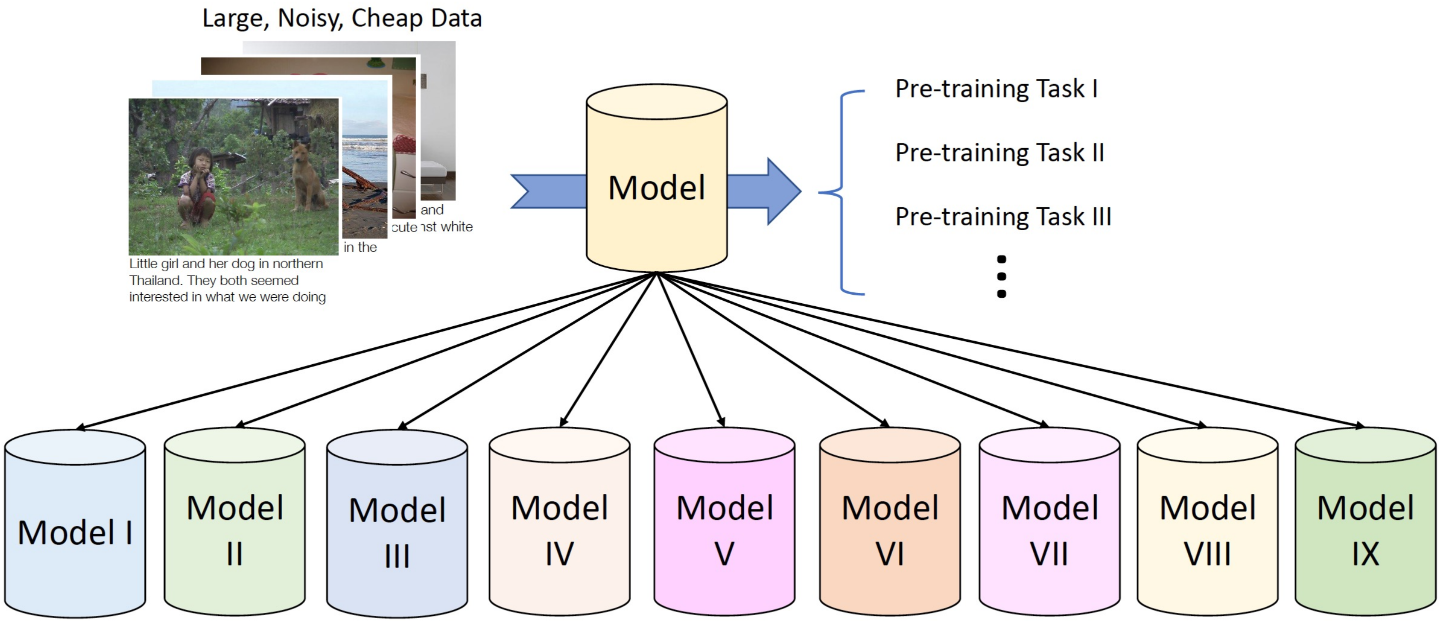


$\mathcal{L}_{\mathrm{M}}$
$\mathcal{L}_{\mathrm{OAC}}$
Vid Proj
Vid Proj
Txt Proj
Txt Proj
CLS
Space-Time Transformer Encoder
Text Encoder
Linear Projection
Tokenizer
X
Mask
Anchor Frame
Road, Water, Tree, Woman, [Random Object]
Visitors take a walk at beautiful Japanese garden

CVPR'21, Microsoft, ClipBert



From retrieval to more tasks



Many downstream tasks/datasets

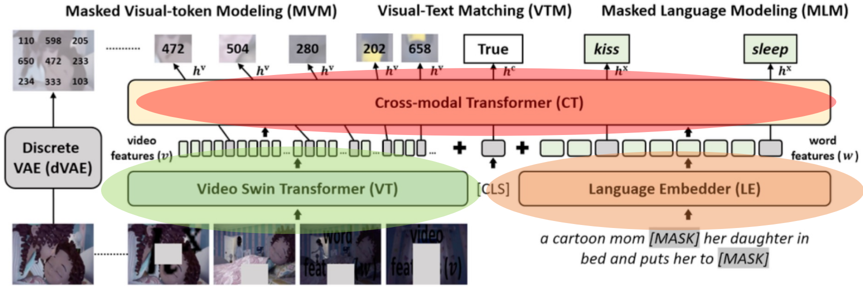
Versatile: transfer to not only many datasets for 1 task, but also to different tasks

[credit to Zhe Gan]

A closer look at these versatile VLP models

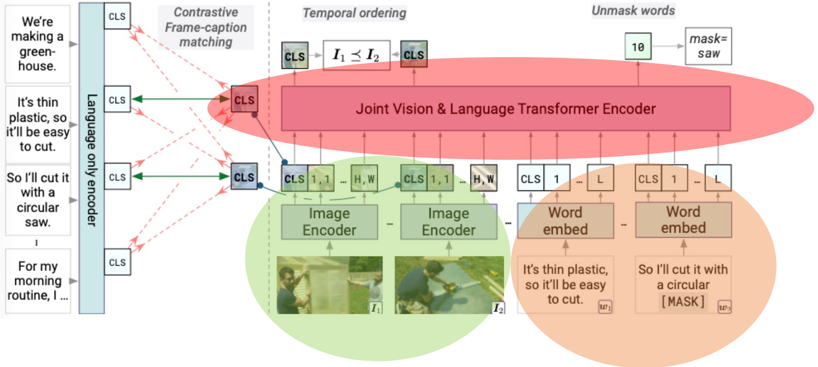
Often have multiple separate components

Arxiv'21, Microsoft, VIOLET



Masked Visual-token Modeling (MVM)
Visual-Text Matching (VTM)
Masked Language Modeling (MLM)
110 598 205
650 472 233
234 333 103
472
504
280
202
658
True
kiss
sleep
Cross-modal Transformer (CT)
Discrete VAE (dVAE)
video features (v)
word features (w)
Video Swin Transformer (VT)
[CLS]
Language Embedder (LE)
a cartoon mom [MASK] her daughter in bed and puts her to [MASK]

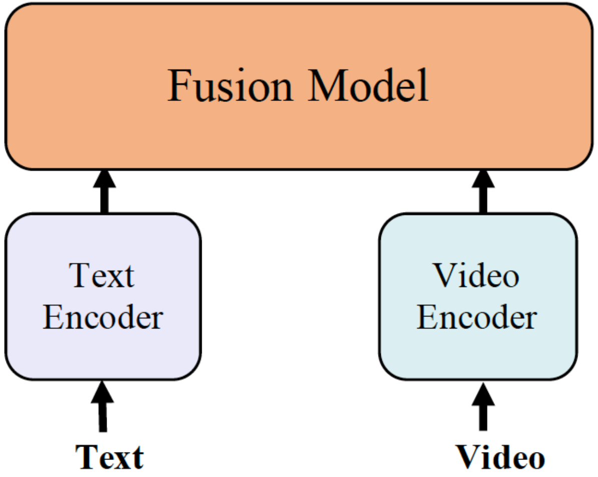
ICML'21, MERLOT



We're making a green-house.
It's thin plastic, so it'll be easy to cut.
So I'll cut it with a circular saw.
For my morning routine, I ...
Language only encoder
CLS
Contrastive Frame-caption matching
Temporal ordering
Unmask words
$I_1 \prec I_2$
10
mask= saw
Joint Vision & Language Transformer Encoder
1,1
H,W
1
L
Image Encoder
Word embed
It's thin plastic, so it'll be easy to cut.
So I'll cut it with a circular [MASK]

A closer look at these versatile VLP models

Often have multiple separate components



Issues:

(1) **Hard to optimize jointly, different components might not be compatible**

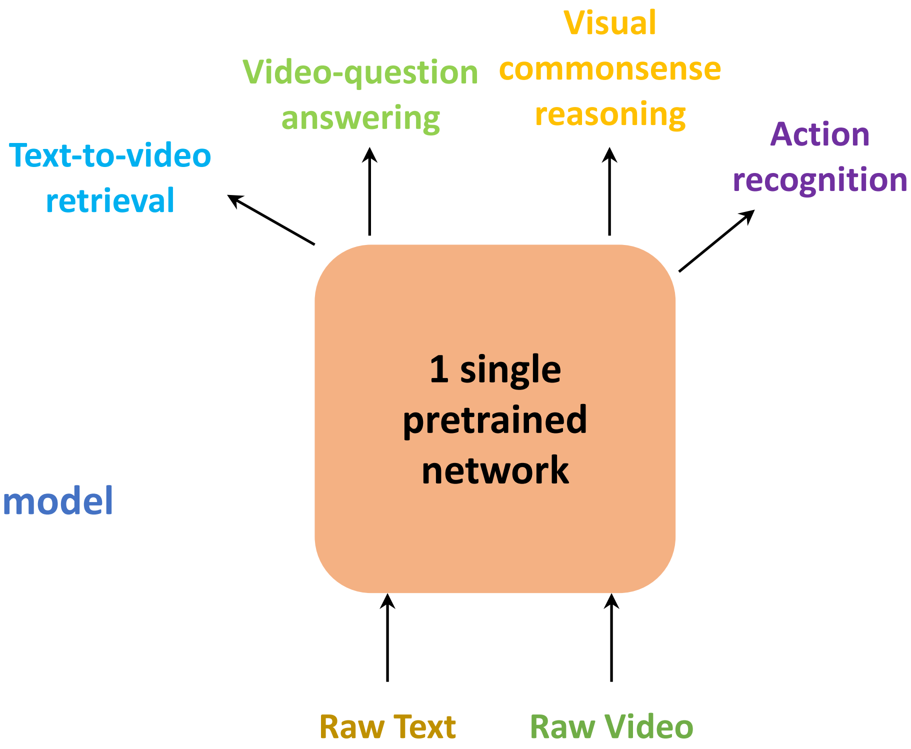
(2) **Redundancy between networks --> share some parameters to save Flops?**

Motivation

Can we have all in one?

(1) All components in one single network

(2) All downstream tasks powered by one pretrained model



All in One: Exploring Unified Video-Language Pre-training

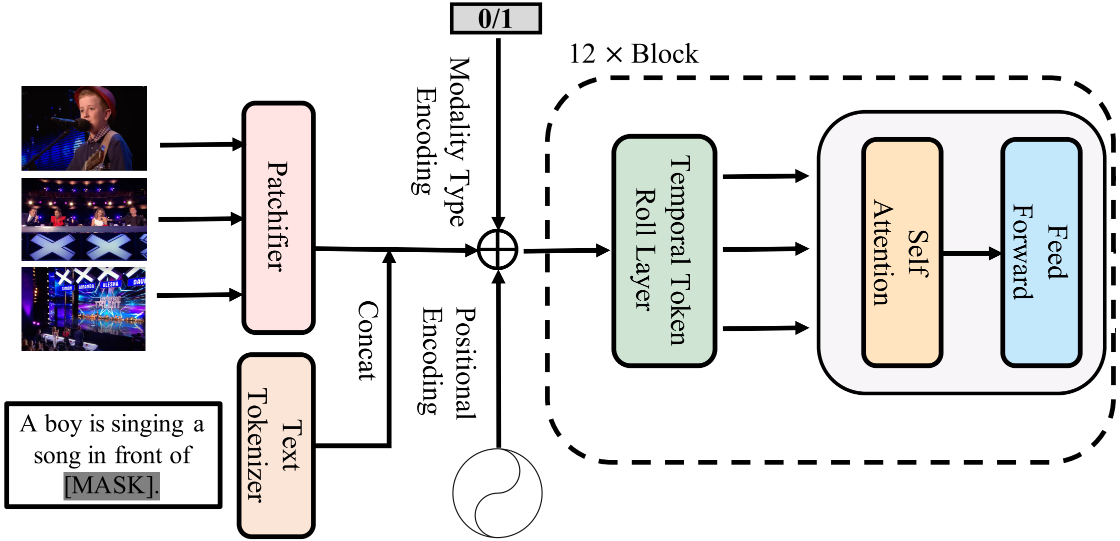
Joint work

w/ Alex Jinpeng Wang

Preprint, 2022.

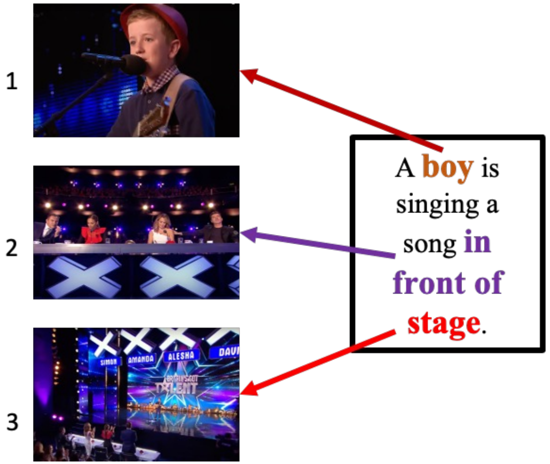
[***https://github.com/showlab/all-in-one***](https://github.com/showlab/all-in-one)

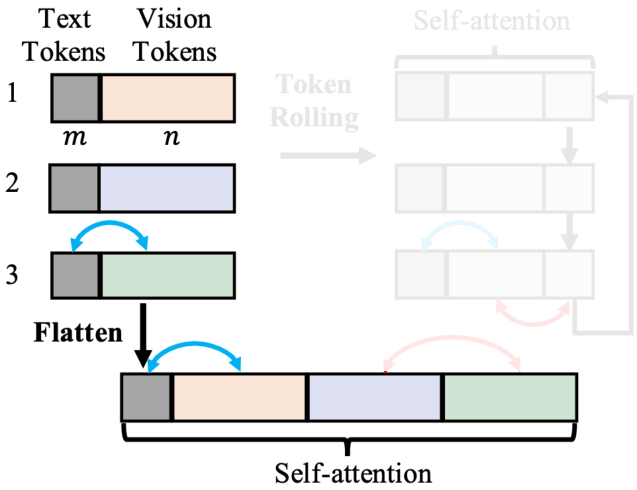
Framework


0/1
12 × Block
Modality Type Encoding
Patchifier
Concat
Positional Encoding
Temporal Token Roll Layer
Self Attention
Feed Forward
Text Tokenizer
A boy is singing a song in front of [MASK].

Temporal Token Rolling Layer

The caption corresponds to multiple frames

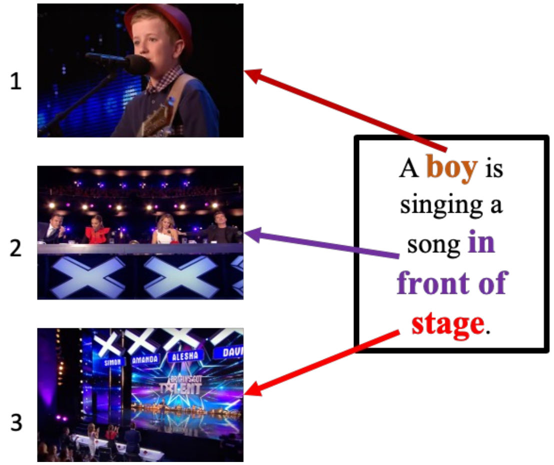


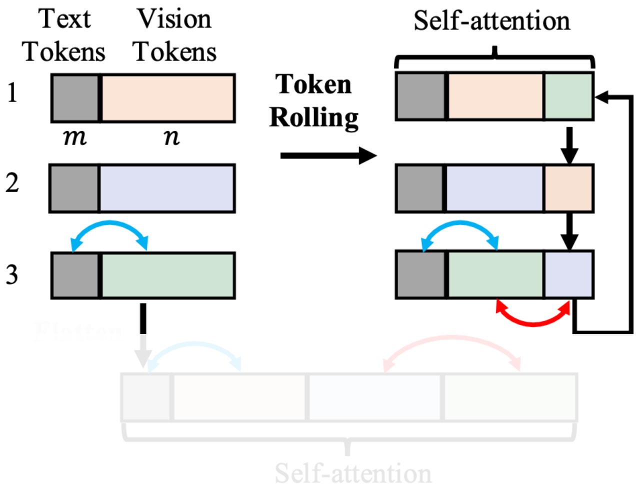


Computational cost is high

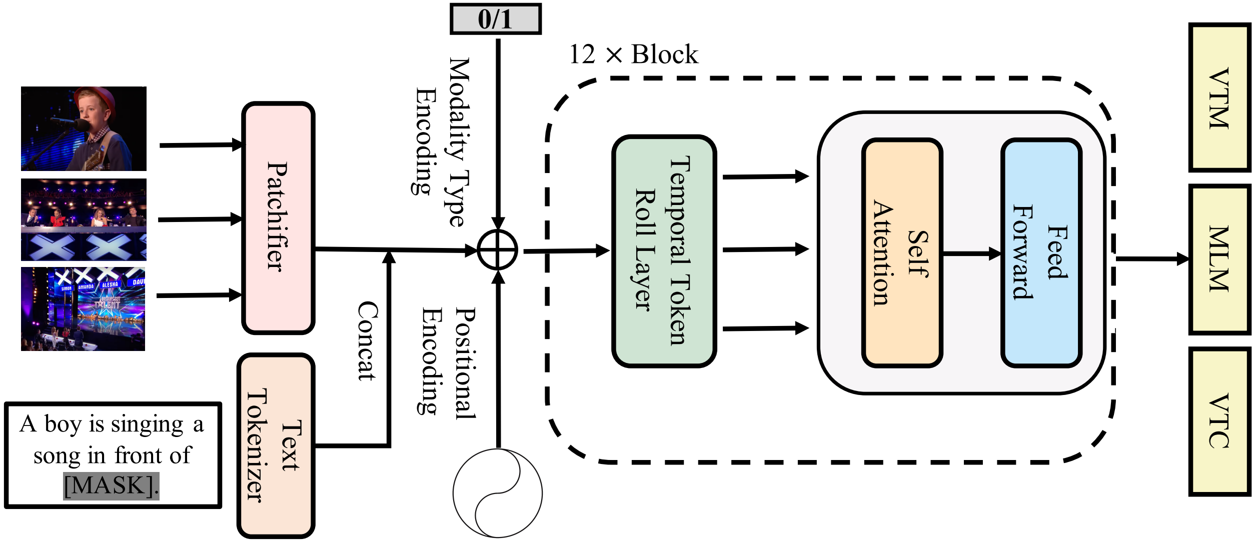
Temporal Token Rolling Layer

- *Model both cross-modality and inter video frames*
- *Parameter-free*





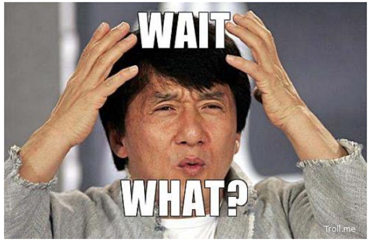
Framework



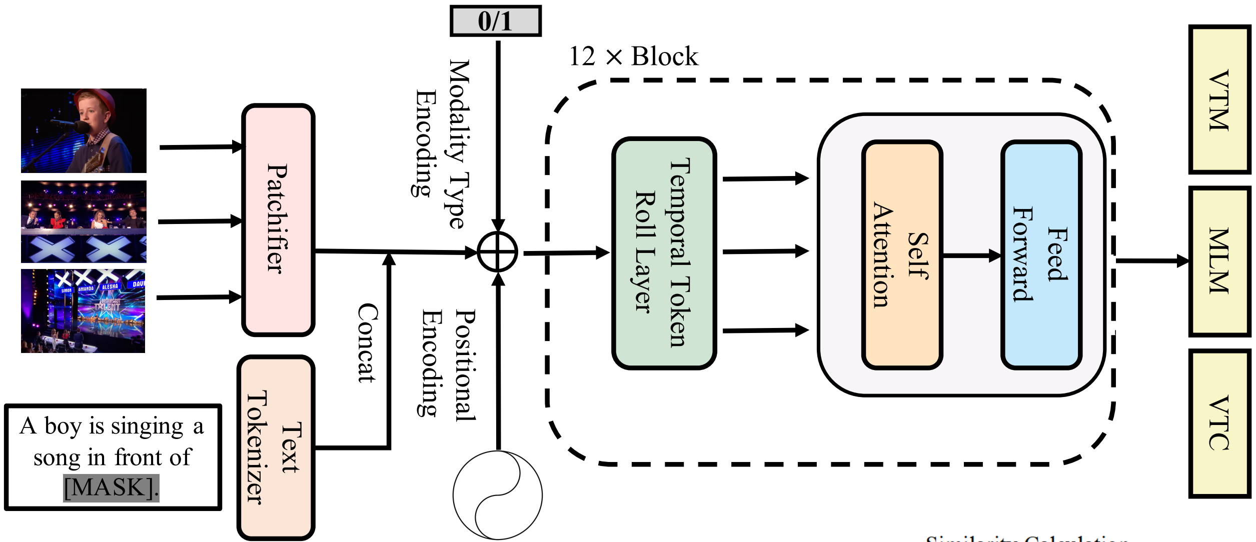
Video-text Matching
(CLS token, binary classifier)

Masked Language Modeling
(text and video)

Video-text
contrastive learning ?



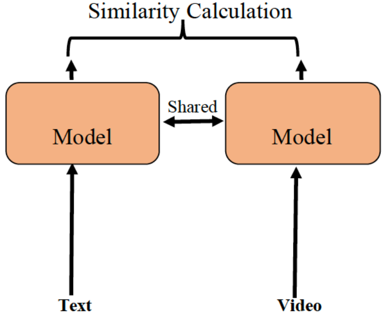
Framework



Video-text Matching (CLS token, binary classifier)

Masked Language Modeling (text and video)

Video-text contrastive learning



Our model can also accept only 1 modality as input.

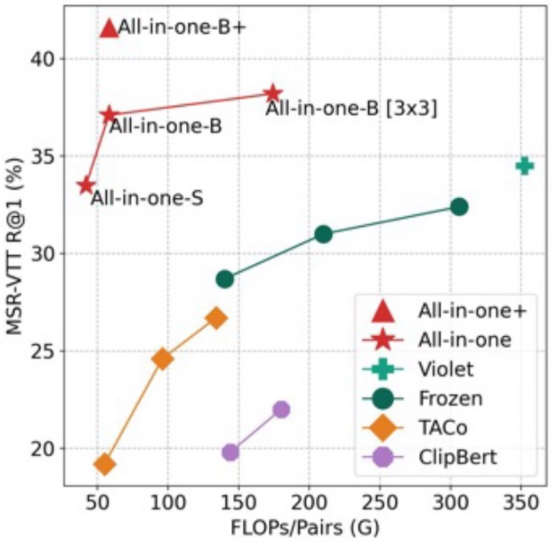
Such design also facilitates the retrieval task which only does linear product between text embedding and video embedding

All-in-one: comparisons with SOTA

All-in-one: comparisons with SOTA

Text-to-video Retrieval

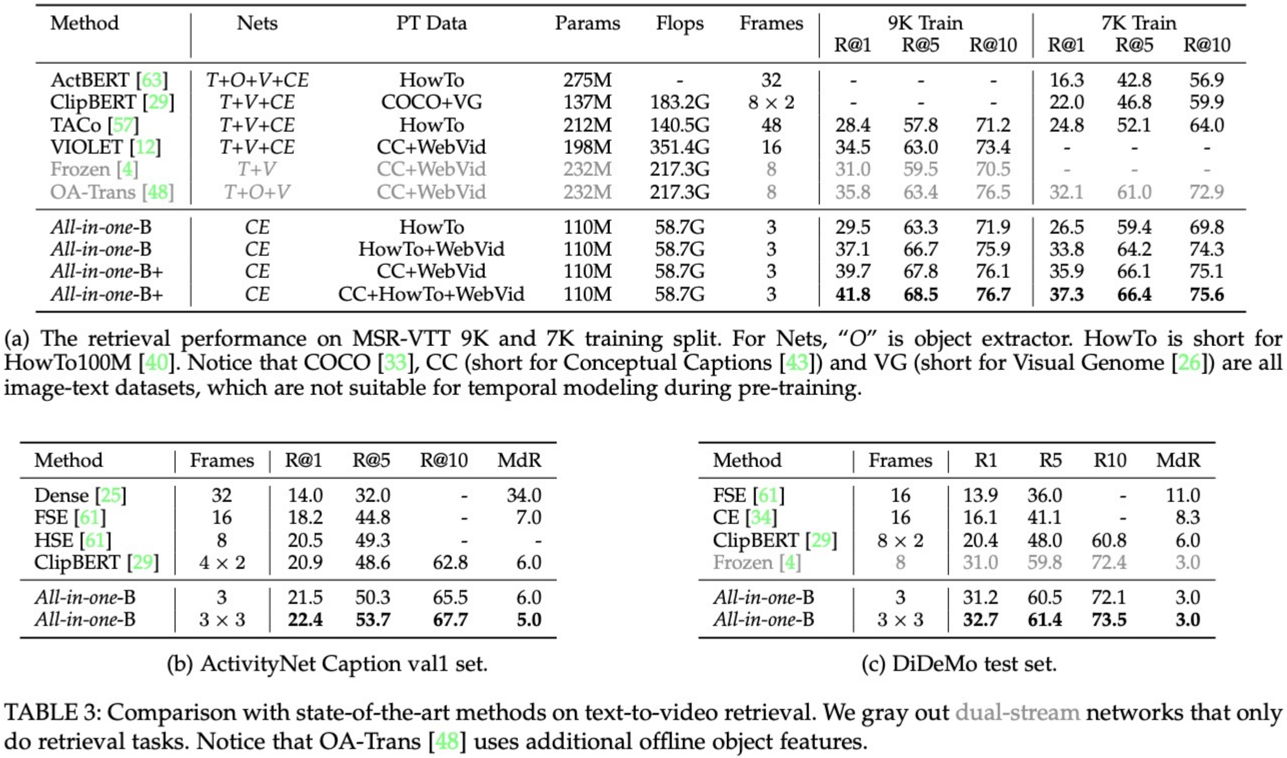
Recall (higher, better)



Efficiency (smaller, better)

All-in-one: comparisons with SOTA

Text-to-video Retrieval on MSR-VTT, ActivityNet Caption, DiDemo



Method	Nets	PT Data	Params	Flops	Frames	9K Train			7K Train		
						R@1	R@5	R@10	R@1	R@5	R@10
ActBERT [63]	*T+O+V+CE*	HowTo	275M	-	32	-	-	-	16.3	42.8	56.9
ClipBERT [29]	*T+V+CE*	COCO+VG	137M	183.2G	8 × 2	-	-	-	22.0	46.8	59.9
TACo [57]	*T+V+CE*	HowTo	212M	140.5G	48	28.4	57.8	71.2	24.8	52.1	64.0
VIOLET [12]	*T+V+CE*	CC+WebVid	198M	351.4G	16	34.5	63.0	73.4	-	-	-
Frozen [4]	*T+V*	CC+WebVid	232M	217.3G	8	31.0	59.5	70.5	-	-	-
OA-Trans [48]	*T+O+V*	CC+WebVid	232M	217.3G	8	35.8	63.4	76.5	32.1	61.0	72.9
All-in-one-B	*CE*	HowTo	110M	58.7G	3	29.5	63.3	71.9	26.5	59.4	69.8
All-in-one-B	*CE*	HowTo+WebVid	110M	58.7G	3	37.1	66.7	75.9	33.8	64.2	74.3
All-in-one-B+	*CE*	CC+WebVid	110M	58.7G	3	39.7	67.8	76.1	35.9	66.1	75.1
All-in-one-B+	*CE*	CC+HowTo+WebVid	110M	58.7G	3	**41.8**	**68.5**	**76.7**	**37.3**	**66.4**	**75.6**

(a) The retrieval performance on MSR-VTT 9K and 7K training split. For Nets, "*O*" is object extractor. HowTo is short for HowTo100M [40]. Notice that COCO [33], CC (short for Conceptual Captions [43]) and VG (short for Visual Genome [26]) are all image-text datasets, which are not suitable for temporal modeling during pre-training.

Method	Frames	R@1	R@5	R@10	MdR
Dense [25]	32	14.0	32.0	-	34.0
FSE [61]	16	18.2	44.8	-	7.0
HSE [61]	8	20.5	49.3	-	-
ClipBERT [29]	4 × 2	20.9	48.6	62.8	6.0
All-in-one-B	3	21.5	50.3	65.5	6.0
All-in-one-B	3 × 3	**22.4**	**53.7**	**67.7**	**5.0**

(b) ActivityNet Caption val1 set.

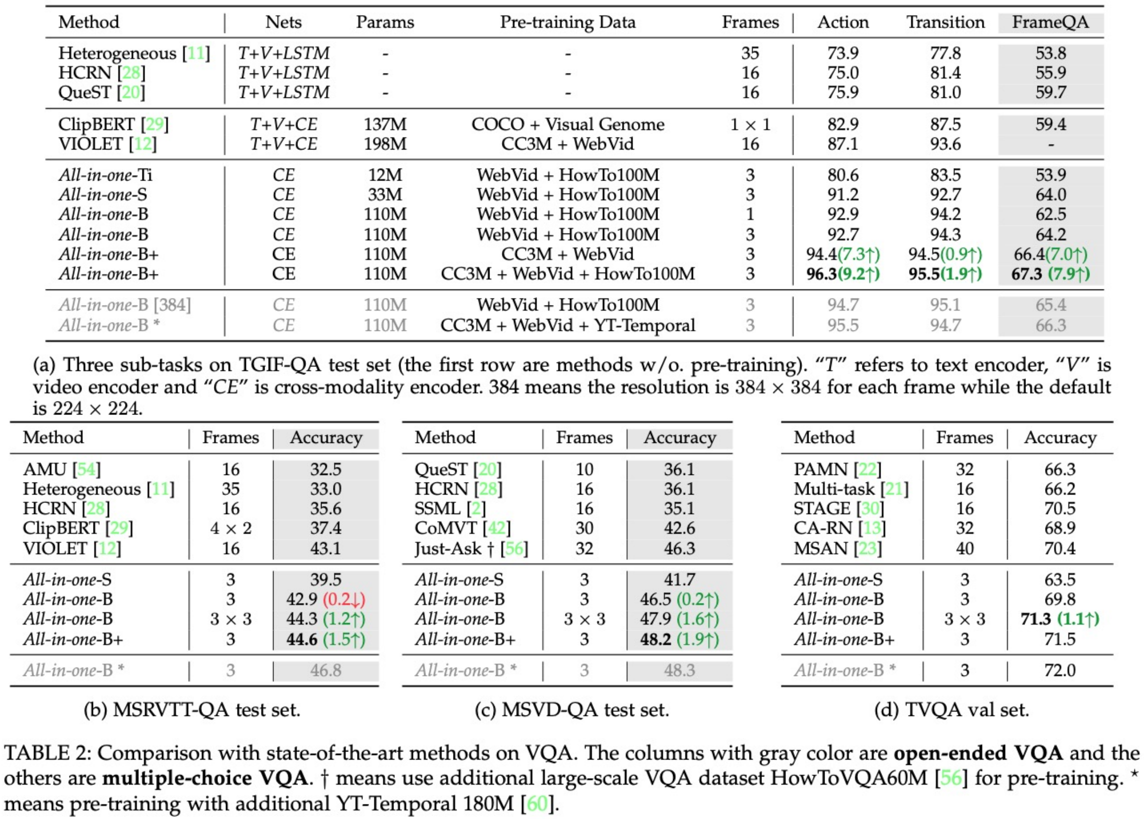
Method	Frames	R1	R5	R10	MdR
FSE [61]	16	13.9	36.0	-	11.0
CE [34]	16	16.1	41.1	-	8.3
ClipBERT [29]	8 × 2	20.4	48.0	60.8	6.0
Frozen [4]	8	31.0	59.8	72.4	3.0
All-in-one-B	3	31.2	60.5	72.1	3.0
All-in-one-B	3 × 3	**32.7**	**61.4**	**73.5**	**3.0**

(c) DiDeMo test set.

TABLE 3: Comparison with state-of-the-art methods on text-to-video retrieval. We gray out dual-stream networks that only do retrieval tasks. Notice that OA-Trans [48] uses additional offline object features.

All-in-one: comparisons with SOTA

Video QA on TGIF-QA, MSRVTT, MSVD-QA, TVQA



Method	Nets	Params	Pre-training Data	Frames	Action	Transition	FrameQA
Heterogeneous [11]	*T+V+LSTM*	-	-	35	73.9	77.8	53.8
HCRN [28]	*T+V+LSTM*	-	-	16	75.0	81.4	55.9
QueST [20]	*T+V+LSTM*	-	-	16	75.9	81.0	59.7
ClipBERT [29]	*T+V+CE*	137M	COCO + Visual Genome	1 × 1	82.9	87.5	59.4
VIOLET [12]	*T+V+CE*	198M	CC3M + WebVid	16	87.1	93.6	-
All-in-one-Ti	*CE*	12M	WebVid + HowTo100M	3	80.6	83.5	53.9
All-in-one-S	*CE*	33M	WebVid + HowTo100M	3	91.2	92.7	64.0
All-in-one-B	*CE*	110M	WebVid + HowTo100M	1	92.9	94.2	62.5
All-in-one-B	*CE*	110M	WebVid + HowTo100M	3	92.7	94.3	64.2
All-in-one-B+	CE	110M	CC3M + WebVid	3	94.4(7.3↑)	94.5(0.9↑)	66.4(7.0↑)
All-in-one-B+	CE	110M	CC3M + WebVid + HowTo100M	3	**96.3(9.2↑)**	**95.5(1.9↑)**	**67.3 (7.9↑)**
All-in-one-B [384]	*CE*	110M	WebVid + HowTo100M	3	94.7	95.1	65.4
All-in-one-B *	*CE*	110M	CC3M + WebVid + YT-Temporal	3	95.5	94.7	66.3

(a) Three sub-tasks on TGIF-QA test set (the first row are methods w/o. pre-training). "*T*" refers to text encoder, "*V*" is video encoder and "*CE*" is cross-modality encoder. 384 means the resolution is 384 × 384 for each frame while the default is 224 × 224.

Method	Frames	Accuracy
AMU [54]	16	32.5
Heterogeneous [11]	35	33.0
HCRN [28]	16	35.6
ClipBERT [29]	4 × 2	37.4
VIOLET [12]	16	43.1
All-in-one-S	3	39.5
All-in-one-B	3	42.9 (0.2↓)
All-in-one-B	3 × 3	44.3 (1.2↑)
All-in-one-B+	3	**44.6** (1.5↑)
All-in-one-B *	3	46.8

(b) MSRVTT-QA test set.

Method	Frames	Accuracy
QueST [20]	10	36.1
HCRN [28]	16	36.1
SSML [2]	16	35.1
CoMVT [42]	30	42.6
Just-Ask † [56]	32	46.3
All-in-one-S	3	41.7
All-in-one-B	3	46.5 (0.2↑)
All-in-one-B	3 × 3	47.9 (1.6↑)
All-in-one-B+	3	**48.2** (1.9↑)
All-in-one-B *	3	48.3

(c) MSVD-QA test set.

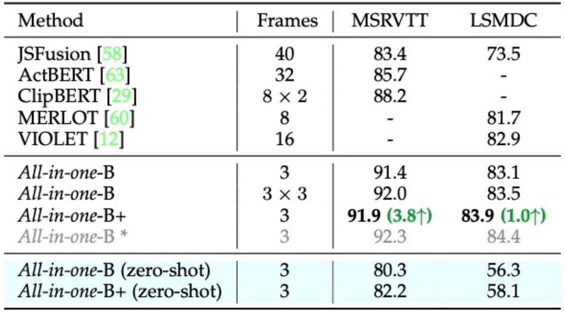
Method	Frames	Accuracy
PAMN [22]	32	66.3
Multi-task [21]	16	66.2
STAGE [30]	16	70.5
CA-RN [13]	32	68.9
MSAN [23]	40	70.4
All-in-one-S	3	63.5
All-in-one-B	3	69.8
All-in-one-B	3 × 3	**71.3 (1.1↑)**
All-in-one-B+	3	71.5
All-in-one-B *	3	72.0

(d) TVQA val set.

TABLE 2: Comparison with state-of-the-art methods on VQA. The columns with gray color are **open-ended VQA** and the others are **multiple-choice VQA**. † means use additional large-scale VQA dataset HowToVQA60M [56] for pre-training. * means pre-training with additional YT-Temporal 180M [60].

All-in-one: comparisons with SOTA

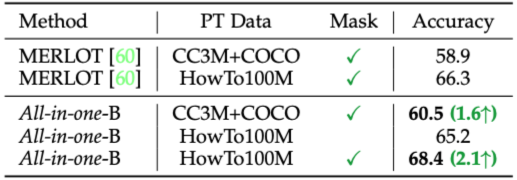
Multiple-choice selection



Method	Frames	MSRVTT	LSMDC
JSFusion [58]	40	83.4	73.5
ActBERT [63]	32	85.7	-
ClipBERT [29]	8 × 2	88.2	-
MERLOT [60]	8	-	81.7
VIOLET [12]	16	-	82.9
All-in-one-B	3	91.4	83.1
All-in-one-B	3 × 3	92.0	83.5
All-in-one-B+	3	**91.9 (3.8↑)**	**83.9 (1.0↑)**
All-in-one-B *	3	92.3	84.4
All-in-one-B (zero-shot)	3	80.3	56.3
All-in-one-B+ (zero-shot)	3	82.2	58.1

TABLE 4: Comparison with state-of-the-art methods on multiple-choice task.

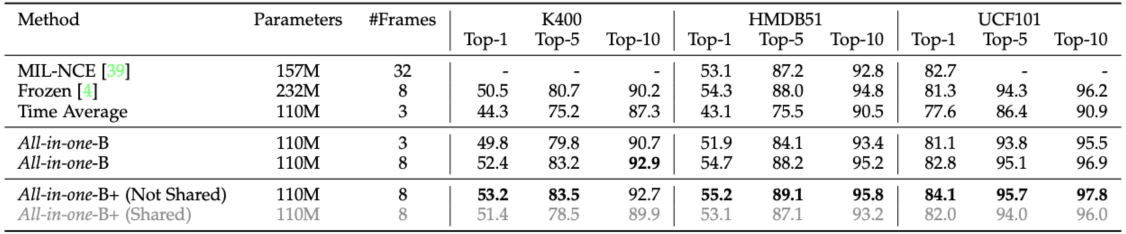
Visual commonsense reasoning



Method	PT Data	Mask	Accuracy
MERLOT [60]	CC3M+COCO	✓	58.9
MERLOT [60]	HowTo100M	✓	66.3
All-in-one-B	CC3M+COCO	✓	**60.5 (1.6↑)**
All-in-one-B	HowTo100M		65.2
All-in-one-B	HowTo100M	✓	**68.4 (2.1↑)**

TABLE 6: The visual commonsense reasoning result with different source of pre-training data.

Action recognition

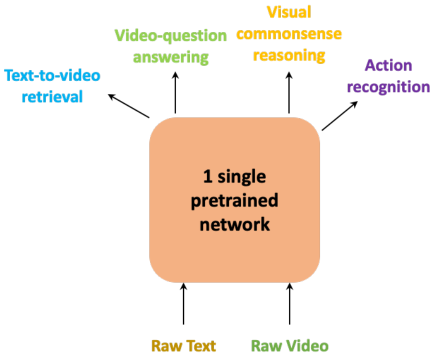


Method	Parameters	#Frames	K400			HMDB51			UCF101		
			Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
MIL-NCE [39]	157M	32	-	-	-	53.1	87.2	92.8	82.7	-	-
Frozen [4]	232M	8	50.5	80.7	90.2	54.3	88.0	94.8	81.3	94.3	96.2
Time Average	110M	3	44.3	75.2	87.3	43.1	75.5	90.5	77.6	86.4	90.9
All-in-one-B	110M	3	49.8	79.8	90.7	51.9	84.1	93.4	81.1	93.8	95.5
All-in-one-B	110M	8	52.4	83.2	**92.9**	54.7	88.2	95.2	82.8	95.1	96.9
All-in-one-B+ (Not Shared)	110M	8	**53.2**	**83.5**	92.7	**55.2**	**89.1**	**95.8**	**84.1**	**95.7**	**97.8**
All-in-one-B+ (Shared)	110M	8	51.4	78.5	89.9	53.1	87.1	93.2	82.0	94.0	96.0

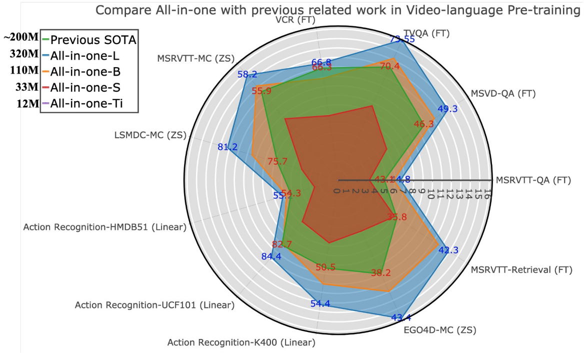
TABLE 9: The linear probe results on action recognition benchmarks over kinetics 400, hmdb51 and UCF101 datasets. Notice that two pre-text heads are not shared for image-text and video-text pairs and the video-text head are used for fine-tuning.

Summary

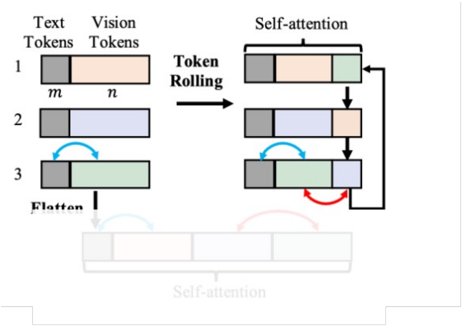
All-in-one, save 50% parameters of SOTA models



SOTA results


Compare All-in-one with previous related work in Video-language Pre-training

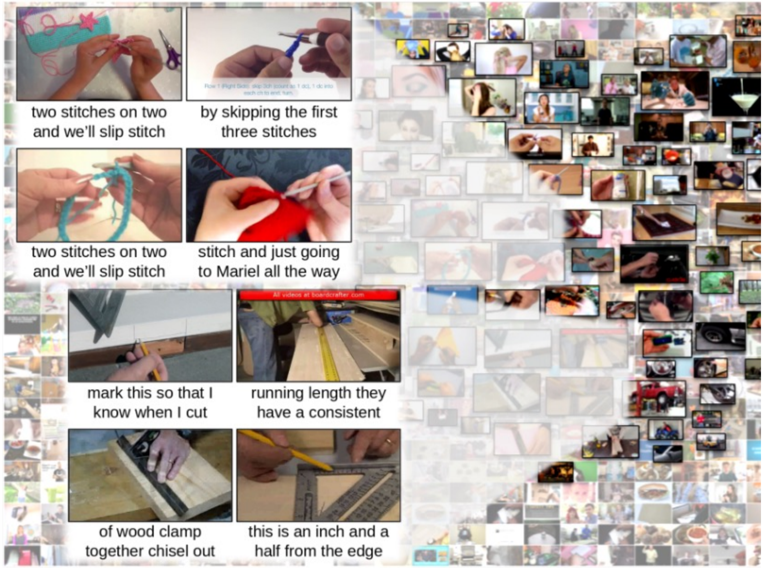
Temporal Token Rolling -- free of parameter



Code & models released

Pretraining videos are of 3rd person view

HowTo100M [ICCV 2019]


two stitches on two and we'll slip stitch
by skipping the first three stitches
two stitches on two and we'll slip stitch
stitch and just going to Mariel all the way
mark this so that I know when I cut
running length they have a consistent
of wood clamp together chisel out
this is an inch and a half from the edge

WebVid 2.5M [ICCV 2021]

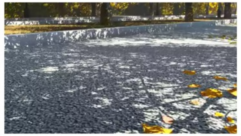
Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking.

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

Billiards, concentrated young woman playing in club.

Cabeza de toro, punta cana/ dominican republic - feb 20, 2020: 4k drone flight over coral reef with manta

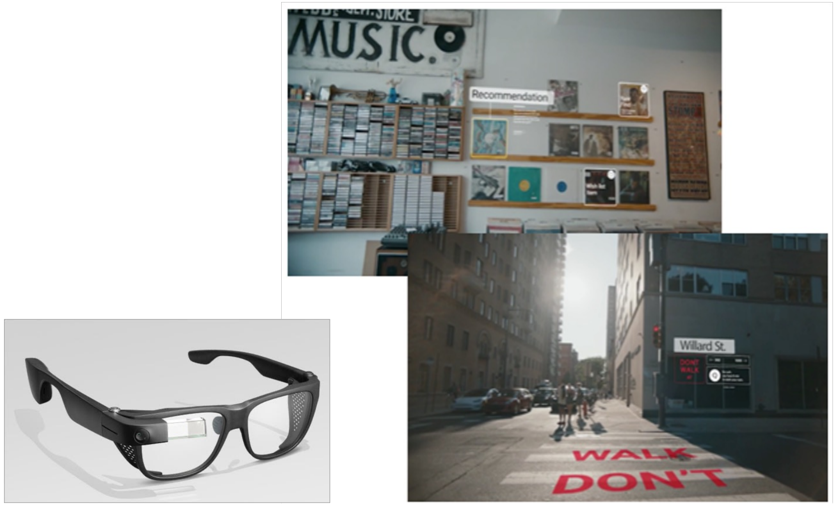
Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing



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

How about egocentric videos?

AR/VR smart glass



Robot learning

[credit to Kristen]

Motivation

Would VLP model pretrained on 3rd person view videos work well for egocentric video?

If not, how can we create an egocentric video-language pretrained (VLP) model?

Egocentric Video-Language Pretraining

Joint work

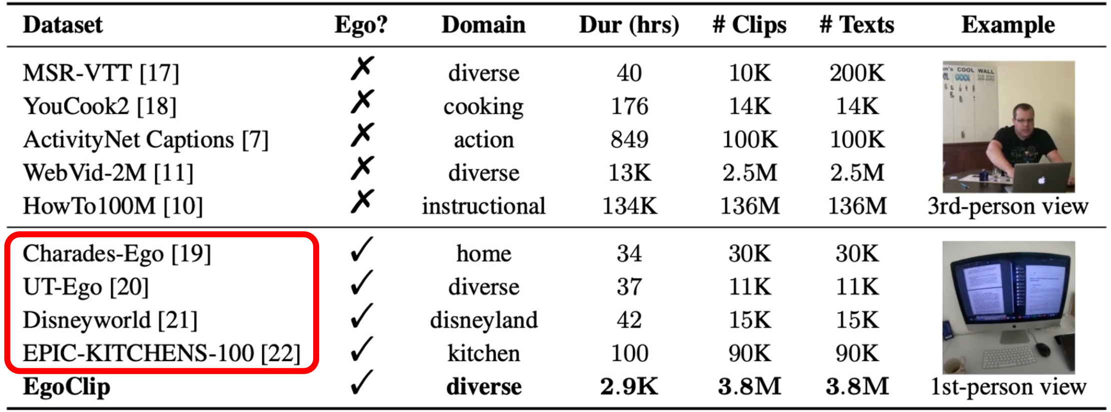
w/ Kevin Qinghong Lin

Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS), 2022.

https://github.com/showlab/EgoVLP

Motivation

- Previous **egocentric datasets** are of **small data scale and domain-specific**, making video-language pre-training impossible.
- **Ego4D** unlocks **Egocentric VLP**!



Dataset	Ego?	Domain	Dur (hrs)	# Clips	# Texts	Example
MSR-VTT [17]	✗	diverse	40	10K	200K	
YouCook2 [18]	✗	cooking	176	14K	14K	
ActivityNet Captions [7]	✗	action	849	100K	100K	
WebVid-2M [11]	✗	diverse	13K	2.5M	2.5M	
HowTo100M [10]	✗	instructional	134K	136M	136M	3rd-person view
Charades-Ego [19]	✓	home	34	30K	30K	
UT-Ego [20]	✓	diverse	37	11K	11K	
Disneyworld [21]	✓	disneyland	42	15K	15K	
EPIC-KITCHENS-100 [22]	✓	kitchen	100	90K	90K	
EgoClip	✓	**diverse**	**2.9K**	**3.8M**	**3.8M**	1st-person view

Table 1: Comparison of our proposed EgoClip pretraining dataset against the mainstream video-language datasets (top) and egocentric datasets (bottom).

Ego4D Data: everyday activity around the world

Data so far:

- 3,600+ hours of video
- ~900 camera wearers
- Geographic diversity
- Occupational diversity
- Unscripted daily life activity
- ~80 real-world scenarios

[https://ego4d-data.org/]

Ego4D for VL Pre-training?

- Research Q1: How to create pre-training **dataset** of video-text pairs?

- Research Q2: How to design pre-training **model**?

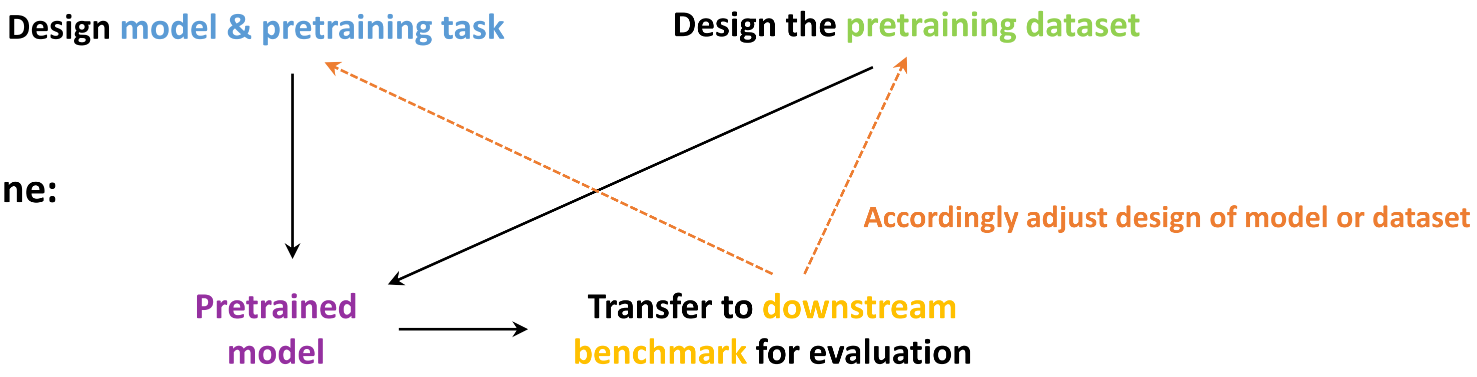
- Research Q3: What benchmark we shall **evaluate** on?

TL;DR

- Create a Large-scale egocentric VL Pre-training set of **3.8M video-text pairs** from Ego4D: **EgoClip**
- Propose an Egocentric-friendly VL pretraining objective: **EgoNCE**
- **Construct a development set for designing Egocentric VL Pre-training: EgoMCQ**

Why need a dev set?

VLP design pipeline:



Design model & pretraining task

Design the pretraining dataset

Pretrained model

Transfer to downstream benchmark for evaluation

Accordingly adjust design of model or dataset

Issue:

when the downstream benchmark is very different from the pretraining task and dataset, the feedback signal may not be accurate

Why need a dev set?

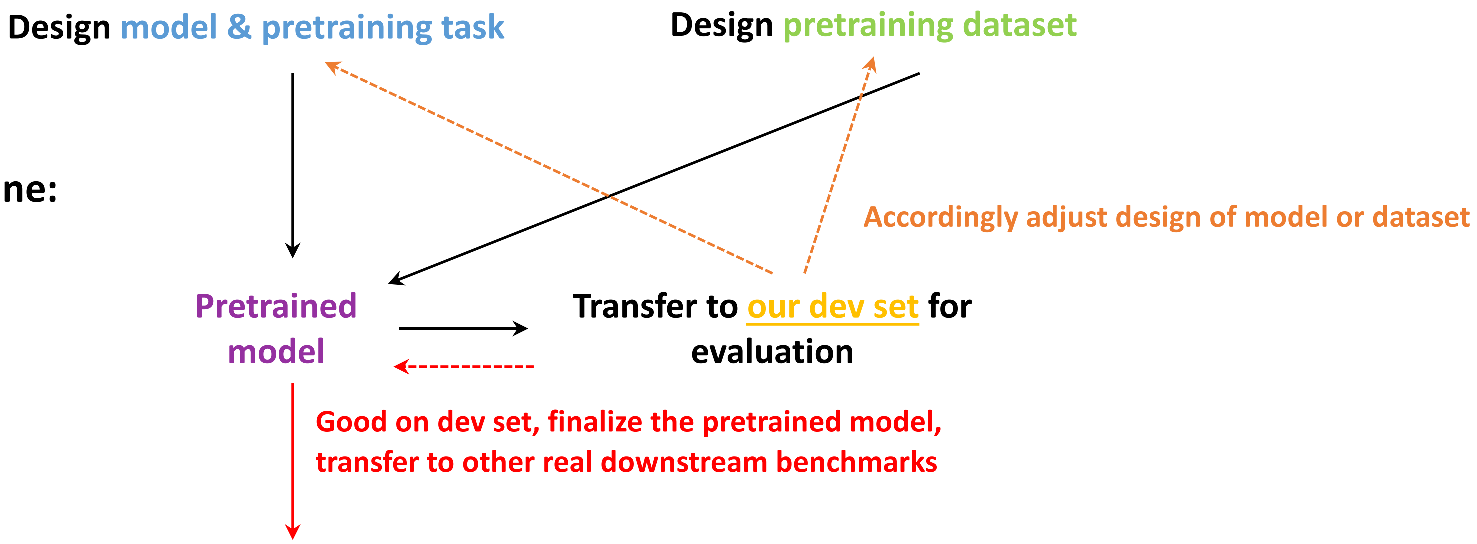
Our Egocentric VLP:

- Pretraining data: in-the-wild
- Pretraining task: video-text matching

Downstream Benchmark	Domain	Task
EPIC-KITCHENs	Kitchen ❌	video-text retrieval ✅
Charades-Ego	Indoor ❌	action recognition ❌
Ego4D benchmarks	In-the-wild ✅	moment localization, object state change detection, etc. ❌
What we'd like to have	In-the-wild ✅	video-text matching ✅

Why need a dev set?

VLP design pipeline:



Design model & pretraining task

Design pretraining dataset

Pretrained model

Transfer to our dev set for evaluation

Accordingly adjust design of model or dataset

Good on dev set, finalize the pretrained model, transfer to other real downstream benchmarks

Downstream Benchmark
EPIC-KITCHENs
Charades-Ego
Ego4D benchmarks

TL;DR

- Create a Large-scale egocentric VL Pre-training set of **3.8M video-text pairs** from Ego4D: **EgoClip**

- Propose an Egocentric-friendly VL pretraining objective: **EgoNCE**

- Construct a development set for designing Egocentric VL Pre-training: **EgoMCQ**

- Significant gains on **5 benchmarks** across **3 datasets:**
  - [EPIC-KITCHENS-100] **Multi-Instance Retrieval**: nDCG (avg) from 53.5% to 59.4%. (+5.9%)
  - [Ego4D Challenges] **Natural Language Query**: R@1 (IoU=0.3) from 5.45% to 10.84%. (+5.4%)
  - [Ego4D Challenges] **Moment Query**: R@1 (IoU=0.3) from 33.45% to 40.43%. (+7.0%)
  - [Ego4D Challenges] **Object State Change Classifcaition**: Acc from 68.7% to 73.9%. (+5.2%)
  - [Charades-Ego] **Action-recognition**: MAP from 30.1% to 32.1%. (+2.0%)

Takeaway

Object-aware Video-language Pre-training for Retrieval. CVPR 2022.

The first to incorporate object region information into video-language pretraining

https://github.com/FingerRec/OA-Transformer

All in One: Exploring Unified Video-Language Pre-training. Preprint, 2022.

All components in 1 single network & all downstream tasks powered by 1 pretrained model, SOTA on 9 datasets across 4 tasks

https://github.com/showlab/all-in-one

Egocentric video-language pretraining. NeurIPS, 2022.

The first to explore egocentric VLP, significant gains on 5 benchmarks across 3 datasets, champion in Ego4D 2022 & Epic-Kitchens 2022 challenges.

https://github.com/showlab/EgoVLP

Thank you!

Q & A

Show

https://sites.google.com/view/showlab