Show

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

All − in − one: Exploring Unified Video-Language Pre-training

Presented by [Alex] Jinpeng Wang

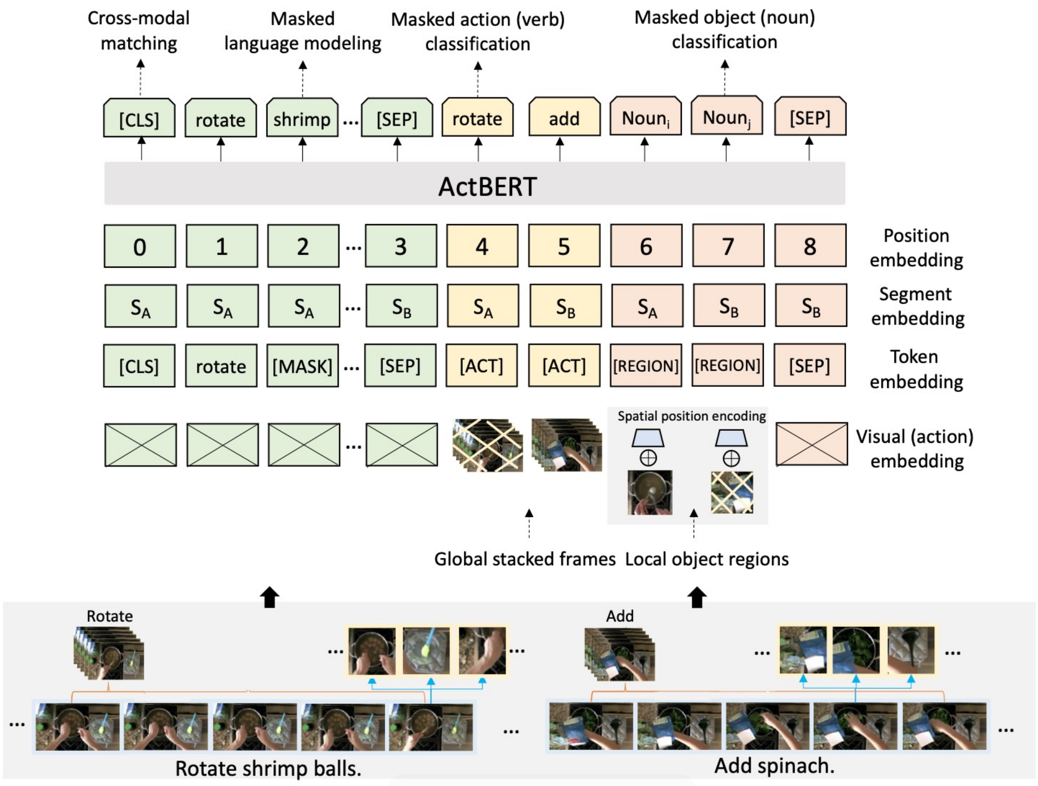
9 20^{th}, 2022

Stage 1: Motivation

Video-language Pretrain Framework

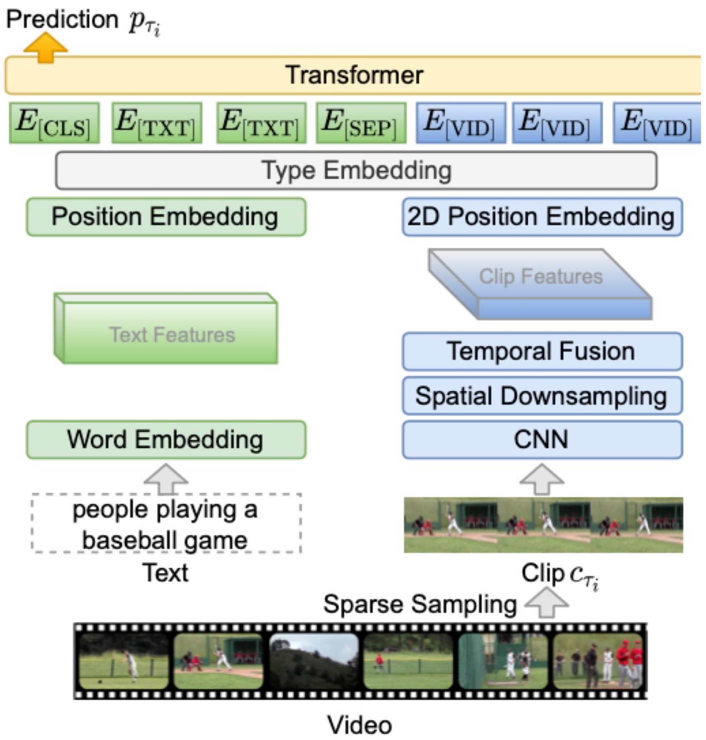
First look at these video-language pretrain framework

3D CNN + Faster RCNN + ActBERT



ActBert, CVPR' 20

2D CNN + BERT + CT

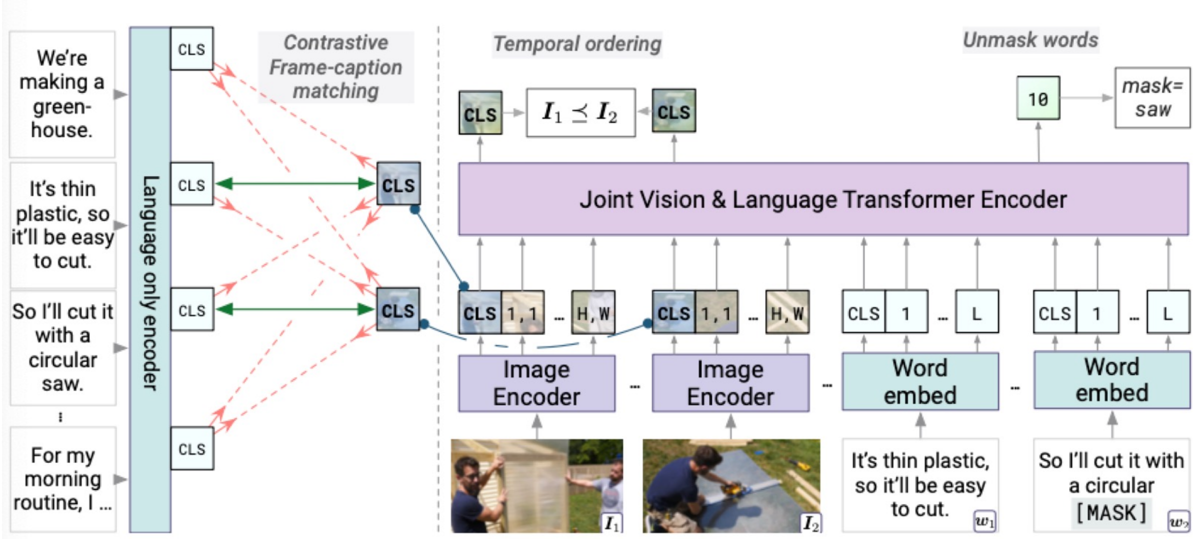


ClipBert, CVPR' 21

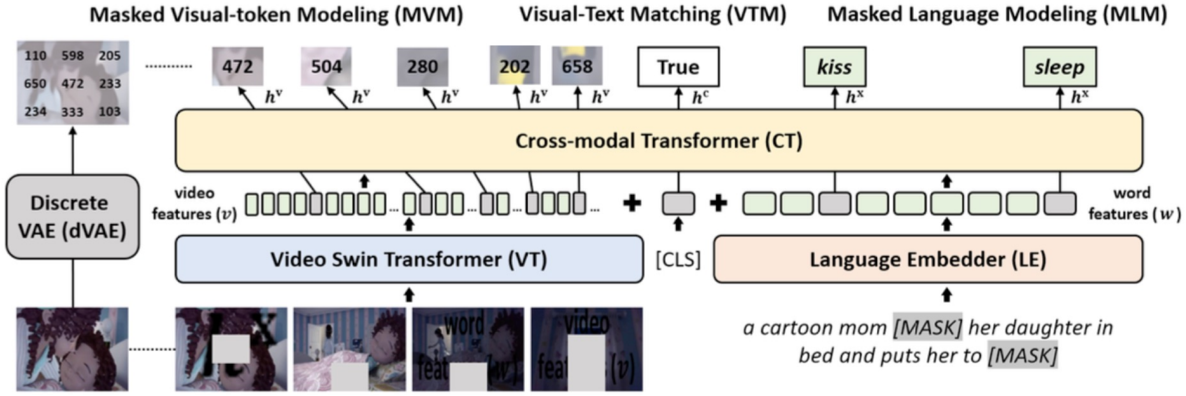
Video-language Pretrain Framework

2D CNN + LE + CT

Video Swin + VAE + LE + CT



MERLOT, ICML' 21



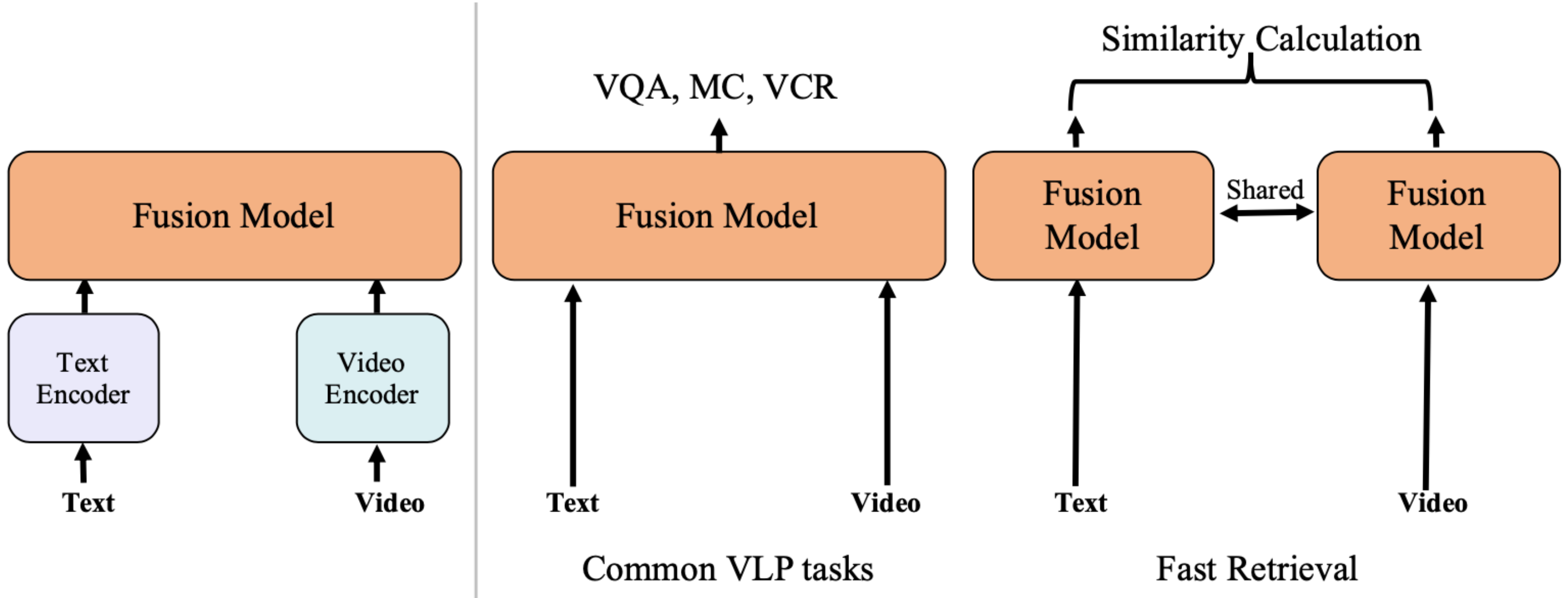
VIOLERT, Arxiv' 21

Video-language Pretrain

- Video-language Pretrain is never an easy thing due to:
  - Optimize at least **3-4 networks**. (hard to train)
  - Large **flops**. (unaffordable)
  - Large-scale data (hundreds of **millions** videos=tens of **billions**-level frames).

Our Motivation: Conduct e2e pretrain with only **1 network with** limited flops & limited frames (sparse sampling).

Perform Modality Interaction



(a). Conventional Pre-training Framework

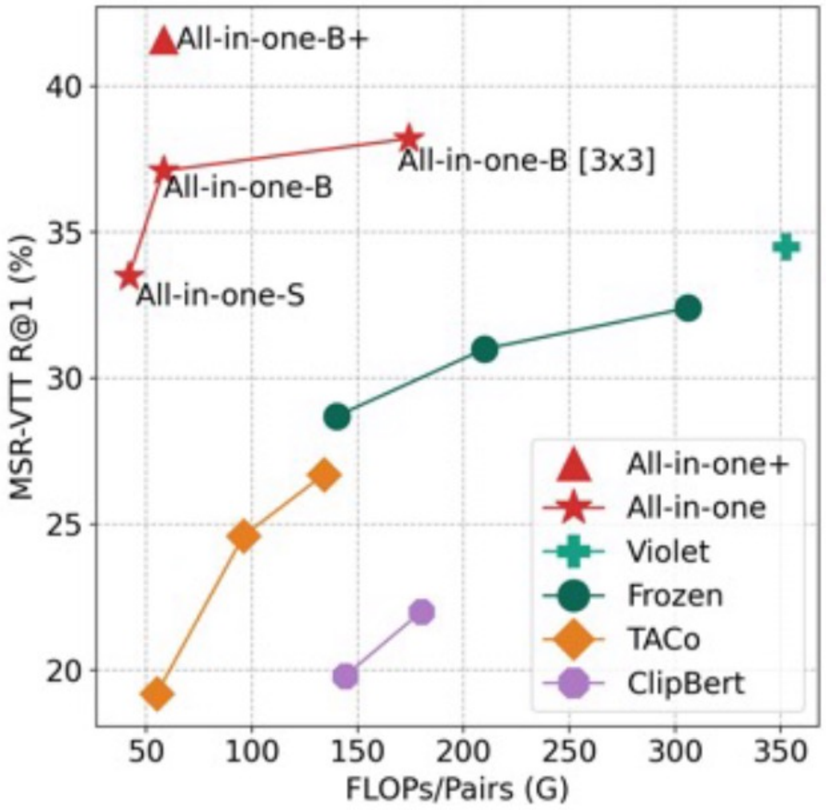
(b). Ours All-in-one Transformer

Get rid of:

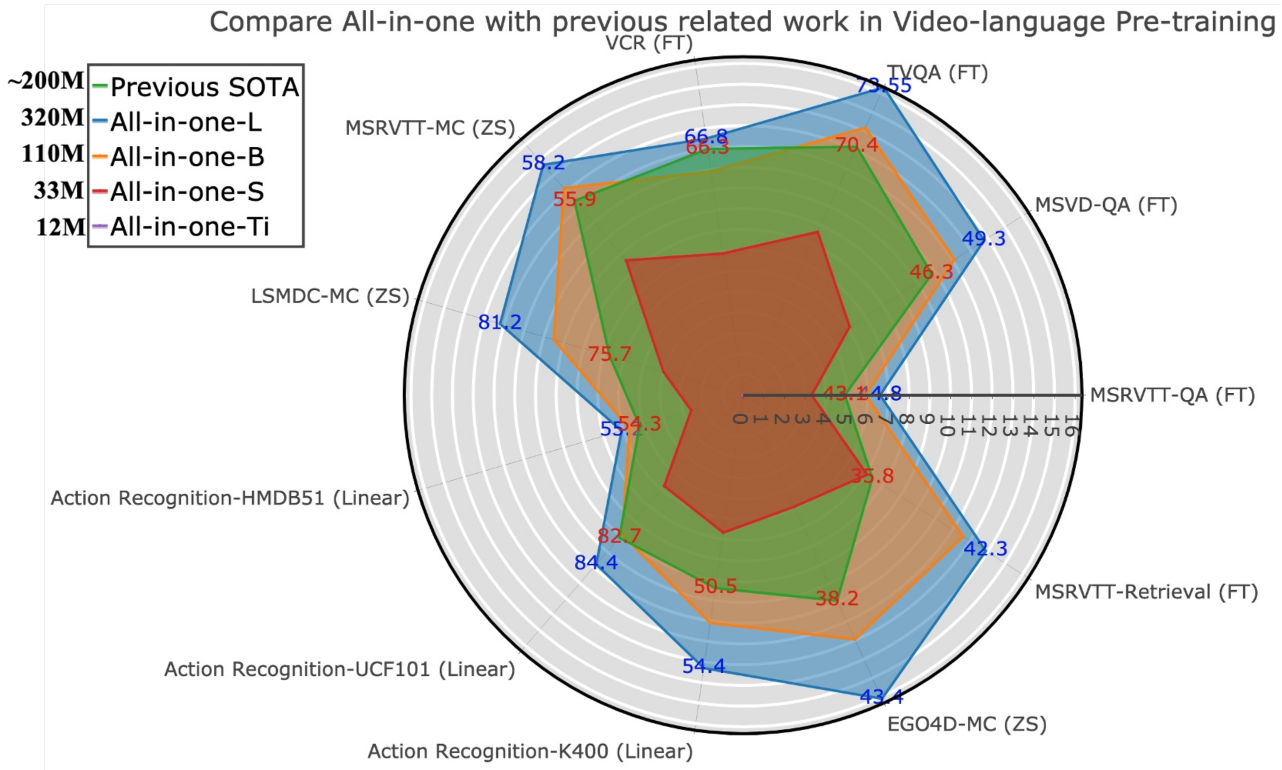
strong and deeper text/vision encoder

train multiple networks with too many hyperparameters

Compare with previous SOTAs



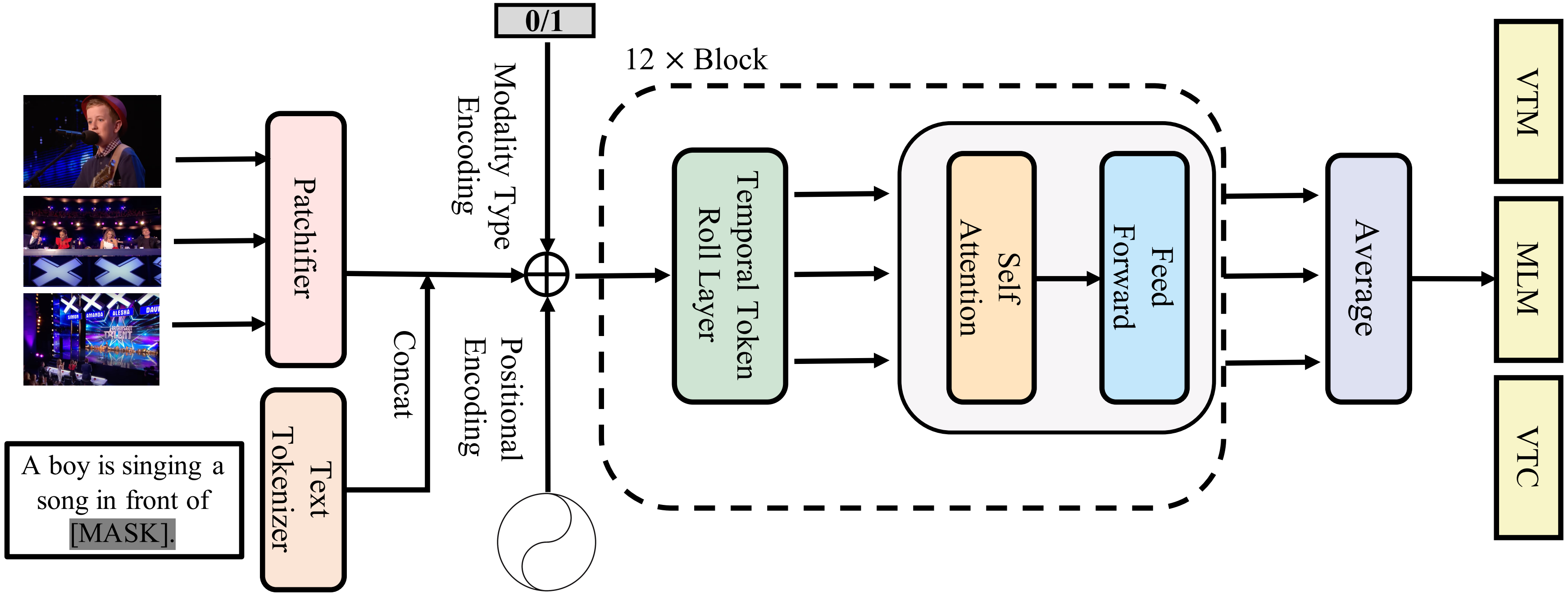
(c). Flops & Performance Comparison



Stage 2: Methodology

All-in-one

- **Shared** self attention/Feed Forward
- **Average** before last head
- Based on VIT



(a). The framework of Temporal Token Roll Transformer (T^2RT)

Fig.2: Model overview. For simplicity, we don't show the normalized layer.

Temporal Token Rolling Layer

Modality interaction require both **short-term** and **long-term reasoning**.

1

2

3

A **boy** is singing a song **in front of stage**.

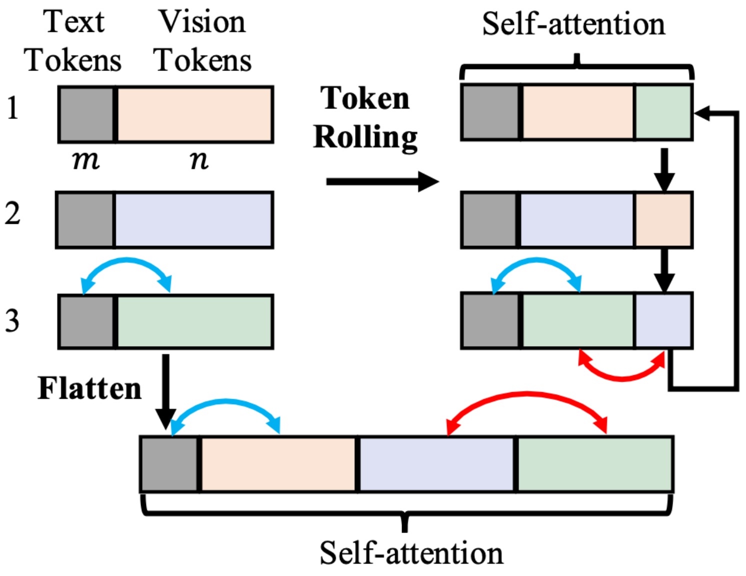


Fig. 3: **The token rolling vs. flatten.** By simply rolling tokens, the computation complex for Self-attention is around one third of Flatten. Not only learns correspondence cross-modality but also inter-frame.

Parameter-free

Stage 3: Experiments

PT & FT Setting

Pretrain

Pretrain 400K steps on 128 NVIDIA A100 GPUs with a batch size of 2,048; Adam.
Warm up 0.1

PT data:
(1). $All - in - one$: Webvid-2.5M, Howto100M (132.5M video-text pairs, 2 days on 128 GPUs)
(2). $All - in - one*$: Webvid-2.5M, Howto100M &YT-Temporal 180M (312.5M video-text pairs, 1 week on 128 GPUs)

Fine-tune

four popular video-and-language tasks:
text-to-video retrieval, video question answering, multiple-choice and visual commensense reasoning across 12 different datasets.
Other tasks include: action recognition, image QA, image-text retrieval.

Video Image Co-training (*All-in-one+*)

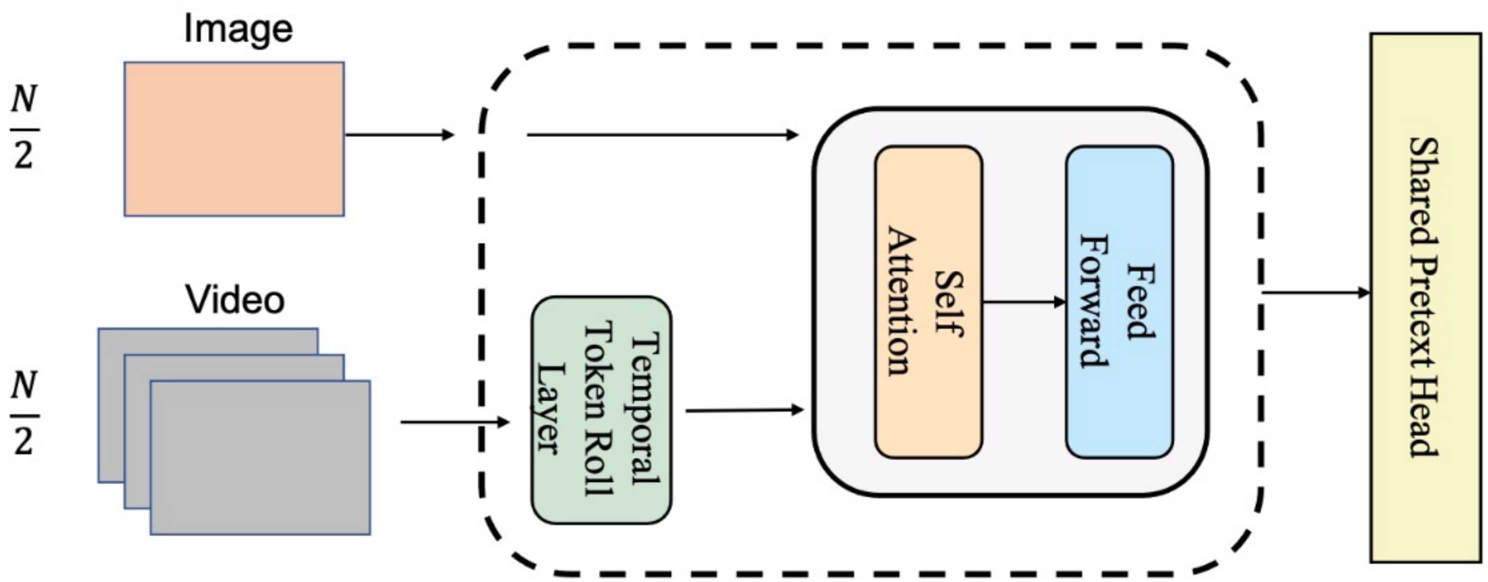


Fig. 5: The image and video co-training pipeline.

Video-Question Answering

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].

Open-ended VQA:
Select 1 answer from N(1600/3100) candidates.
Multiple Choices VQA:
Question & Answers are both sentences.

Video-text Retrieval

Method	Nets	PT Data	Params	Flops	Frames	9K Train R@1	9K Train R@5	9K Train R@10	7K Train R@1	7K Train R@5	7K Train 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.

Multiple-Choice & Visual Commonsense Reasoning

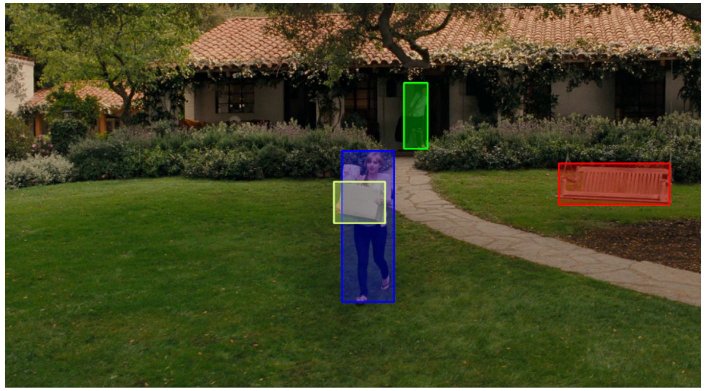
Method	Frames	MSRVTT	LSMDC
JSFusion [58]	40	83.4	73.5
ActBERT [63]	32	85.7	-
ClipBERT [29]	8×2	88.2	-
MERLOT [60]	8	-	81.7
VIOLET [12]	16	-	82.9
All-in-one-B	3	91.4	83.1
All-in-one-B	3×3	92.0	83.5
All-in-one-B+	3	**91.9** (3.8↑)	**83.9** (1.0↑)
All-in-one-B *	3	92.3	84.4
All-in-one-B (zero-shot)	3	80.3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

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.

Color Mask:



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.

Cloze evaluation

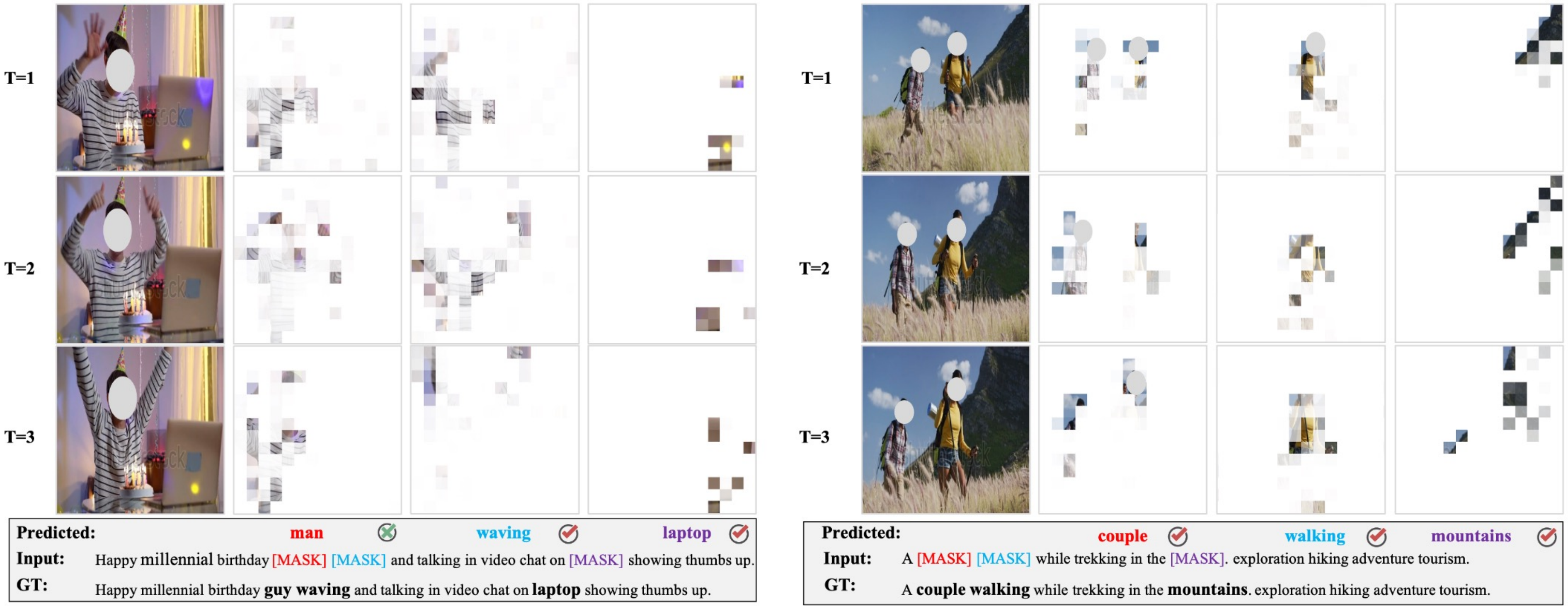


Fig. 8: ***Cloze* evaluation:** Given a video and its paired masked text, the model is asked to *fill the masked words and show its corresponding high attention patch for this masked word*. These samples are sampled from the validation set of Webvid [4].

Image Video Co-training Visualization

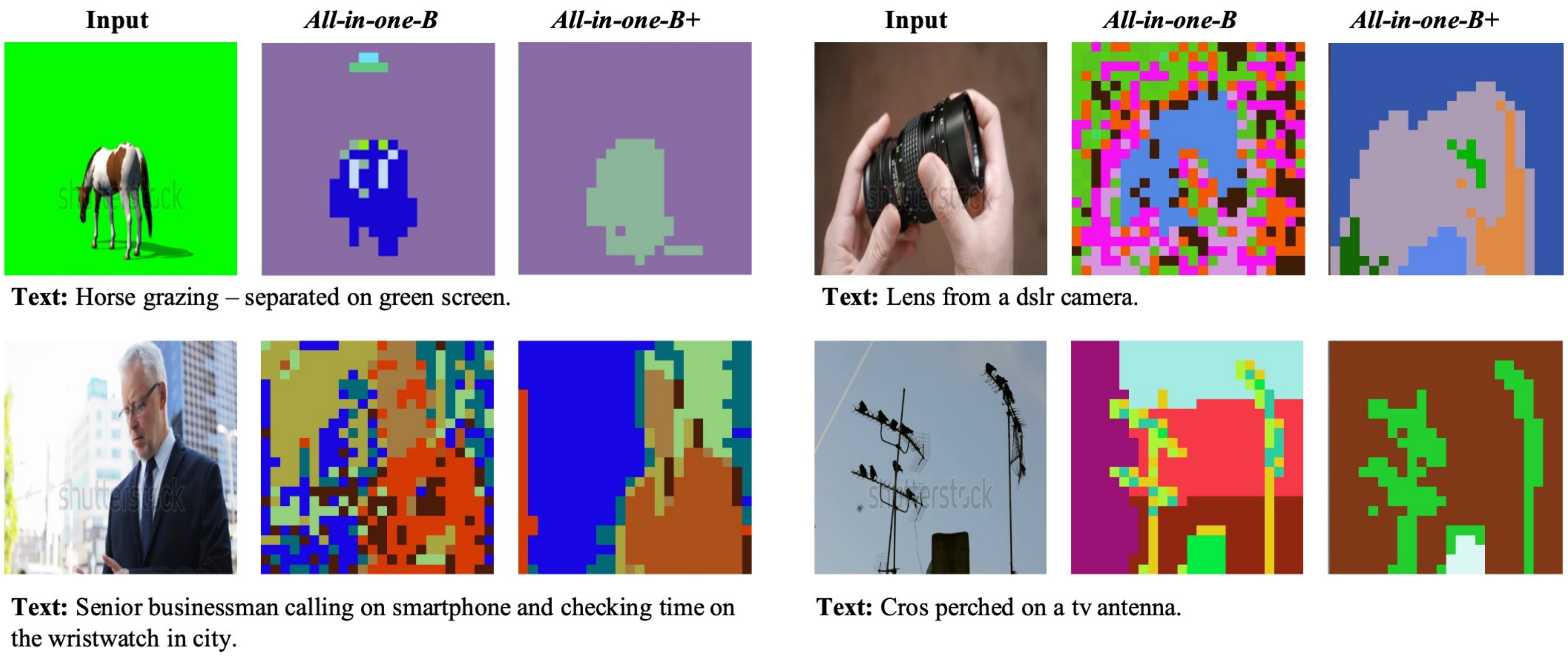


Fig. 12: The token cluster visualization result. By co-training with well-annotated image dataset, the *All-in-one* learns better cluster.

Contribution

- 1. The **first** e2e early-fusion (independent encoder free) one-stream framework in video-language pretrain, with almost 50% parameters and 30% flops among existing pretrain framework.
- 2. A novel parameter-free Temporal Token Rolling for temporal alignment.
- 3. With 3 frames input, T2RT leading to competitive even better results previous sota results (16+frames) on 4 benchmarks.
- Release a simplest codebase for video-language pre-training.

Code

Thanks & QA!

Show

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