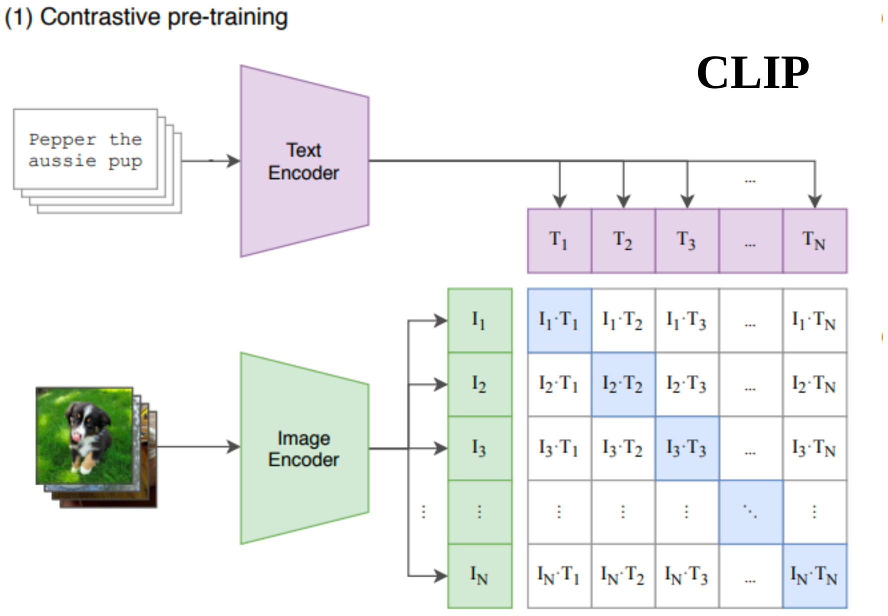
Revisiting Temporal Modeling for CLIP-based Image-to-Video KnowledgeTransferring

Ruyang Liu [1*], Jingjia Huang [2*], Ge Li [1], Jiashi Feng [2], Xinglong Wu [2], Thomas Li [1]

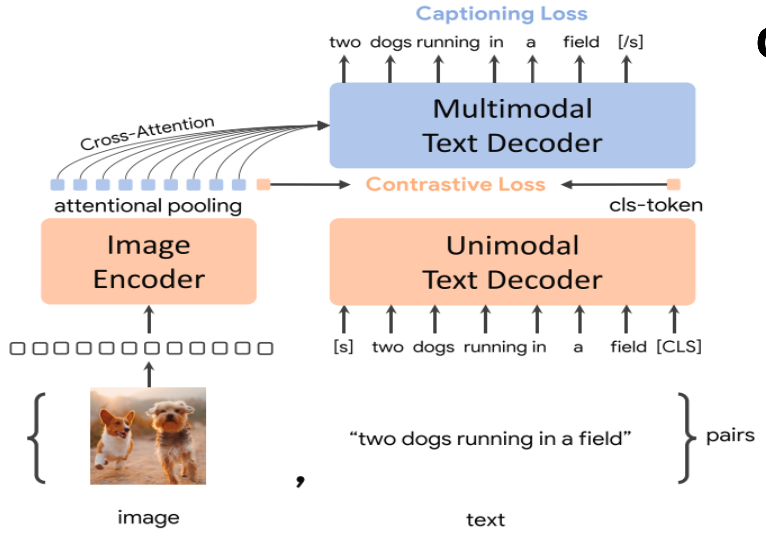
1 Peking University, School of Electronic and Computer Engineering
2 ByteDance Inc, Intelligent Creation Team

Background

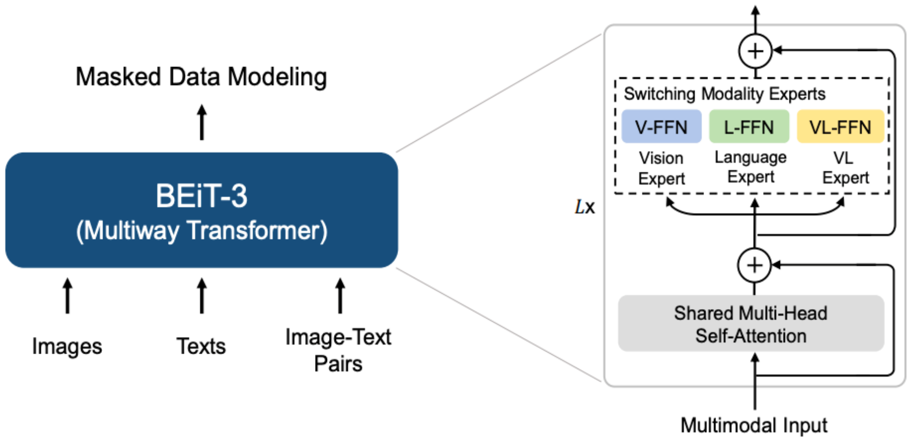
CLIP

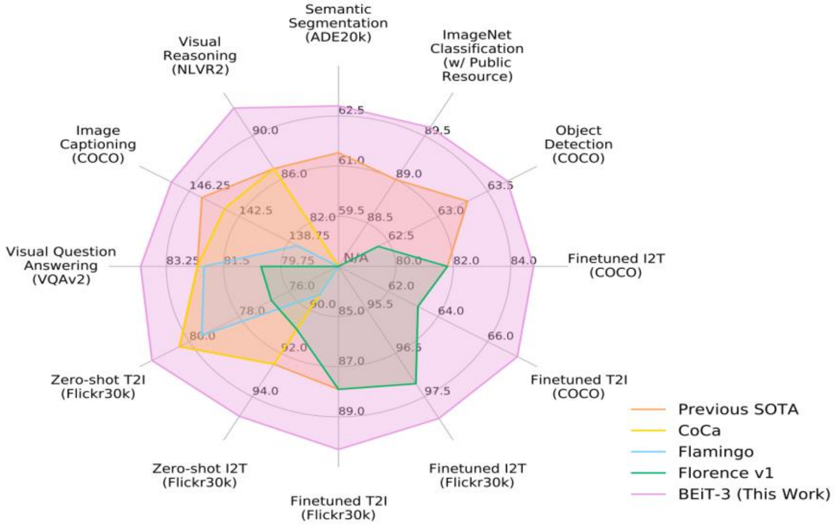


CoCa

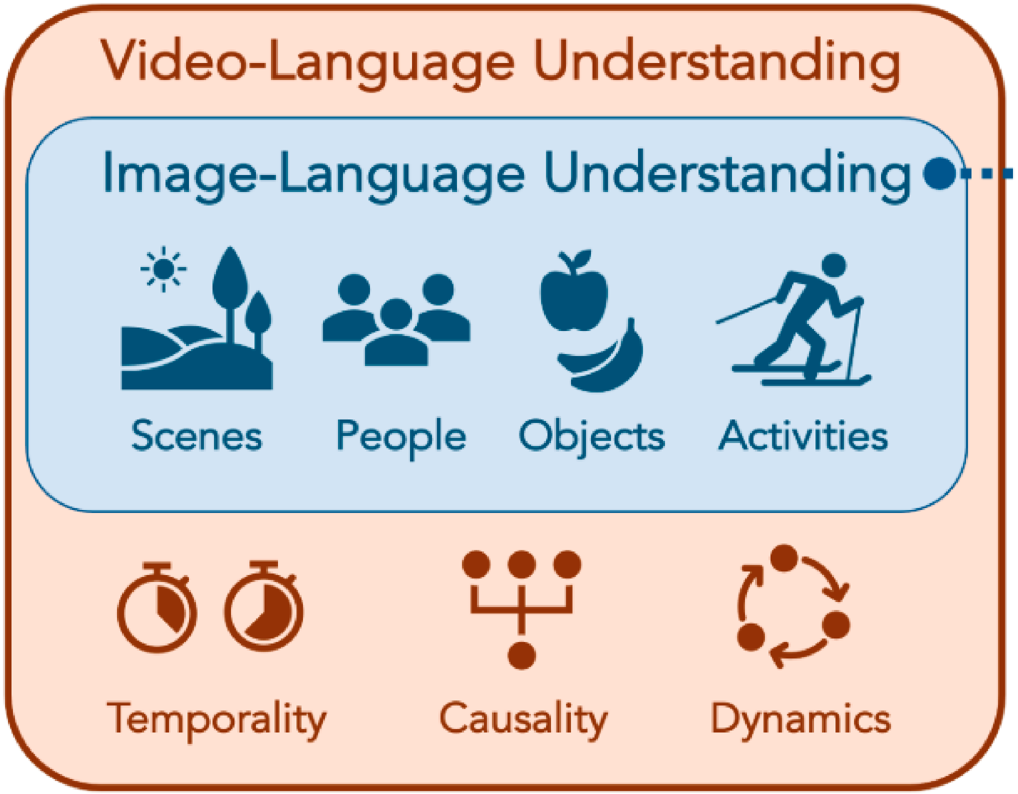


Beit-V3





Background



Video-text pretrained model?

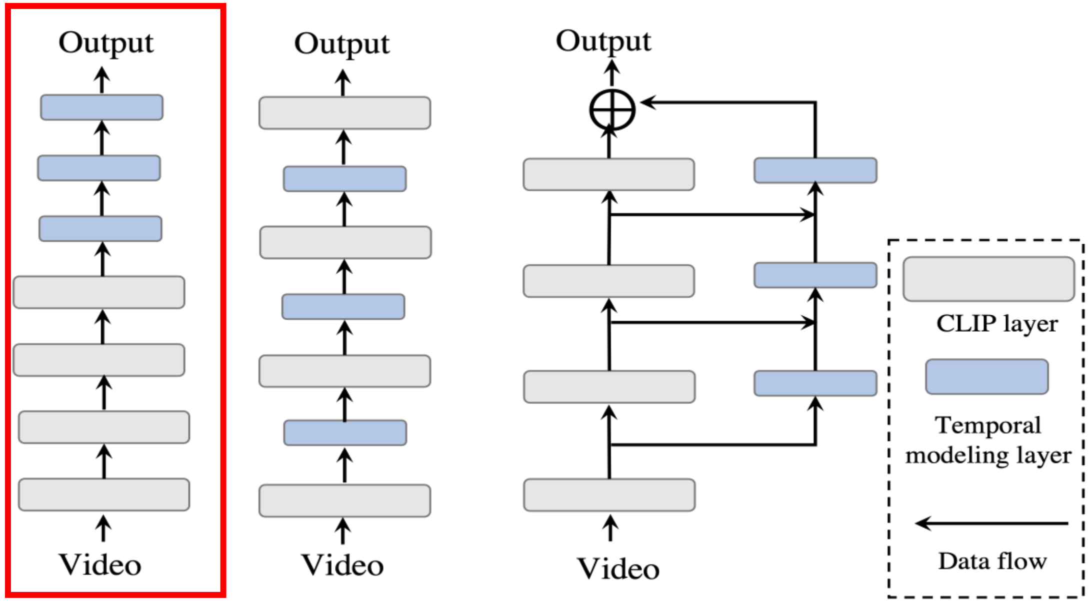
- hard to collect video-text as diverse and large in scale as image-text data
- computational consumption

Image-Text pretrained model

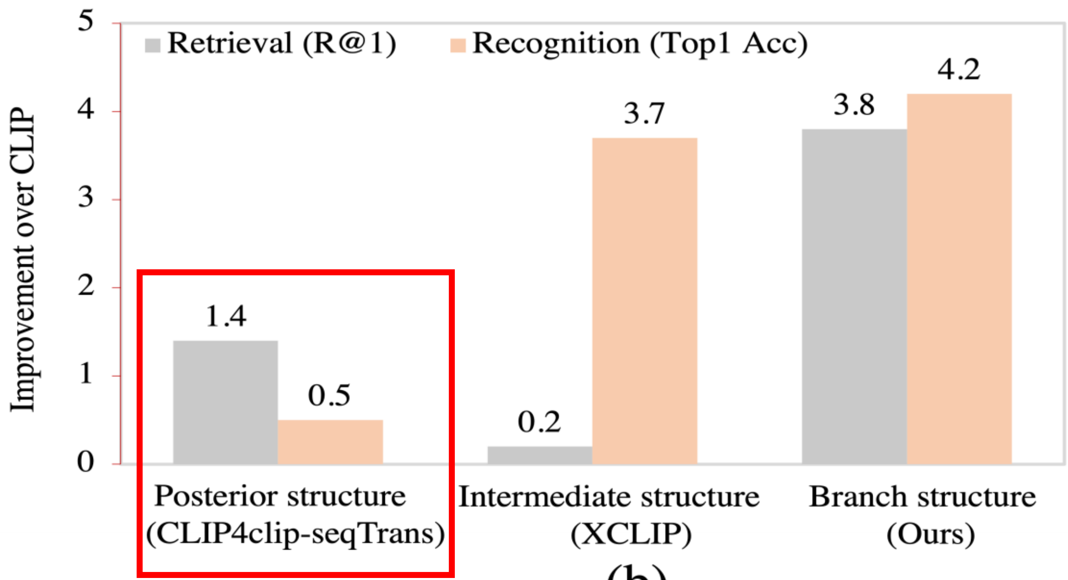
Video domain

- video-text retrieval
- video recognition
- …

Introduction



(a)



(b)

Baseline=CLIP with mean pooling

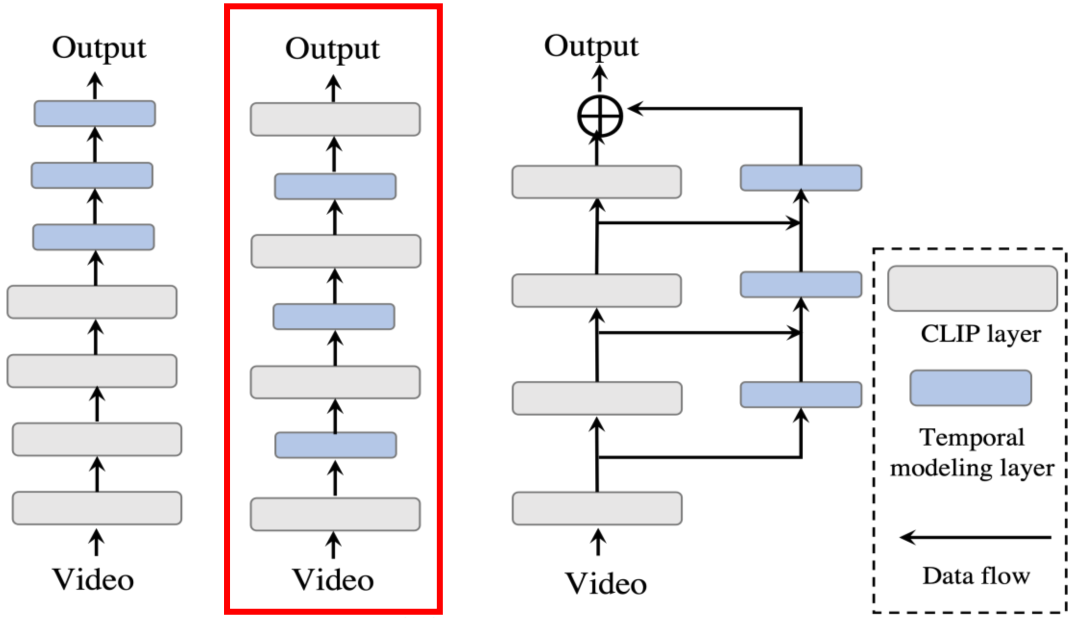
Posterior structure：

- late fusion upon highly semantic embeddings
- downstream video-text retrieval task
- benefit to transfer well aligned visual-language representation (i.e., high-level knowledge)

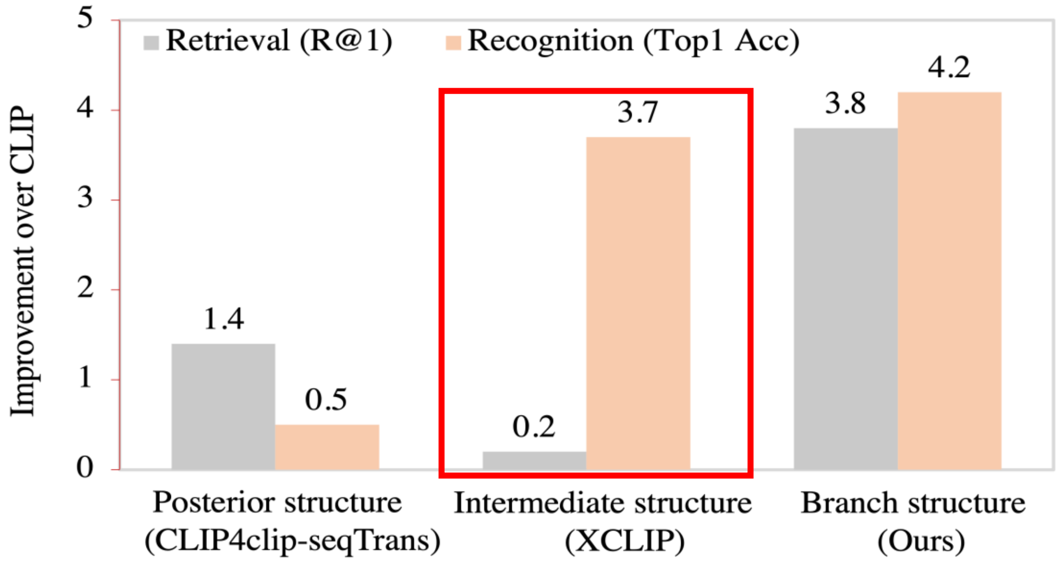
Cons:

- a sub-optimal temporal modeling strategy that hardly capture the spatial-temporal visual patterns

Introduction



(a)



(b)

Baseline=CLIP with mean pooling

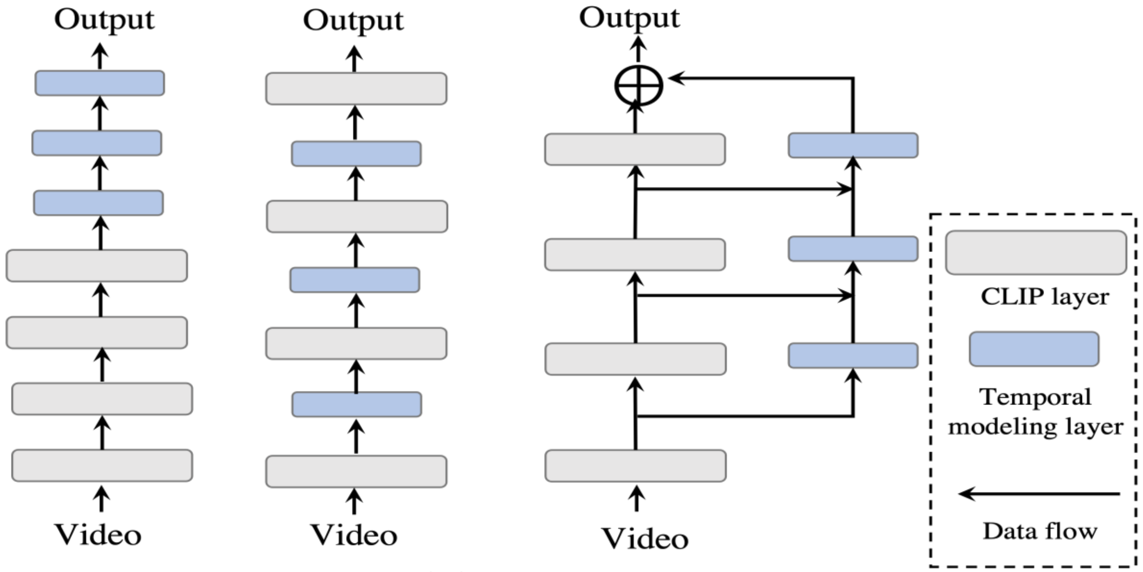
Intermediate structure：

- Intermediate fusion
- downstream video recognition task
- benefit from the pretrained visual patterns (i.e., low-level knowledge) to strengthen spatial-temporal modeling capability of CLIP

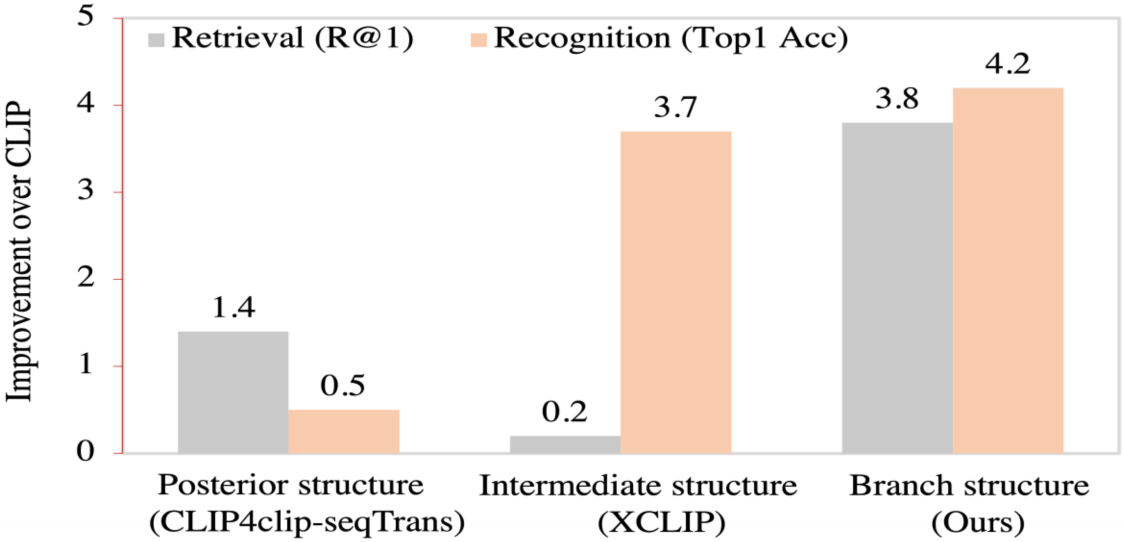
Cons:

- impact the pretrained high-level knowledge which brings about trivial improvement to video-text retrieval

Introduction



(a)



(b)

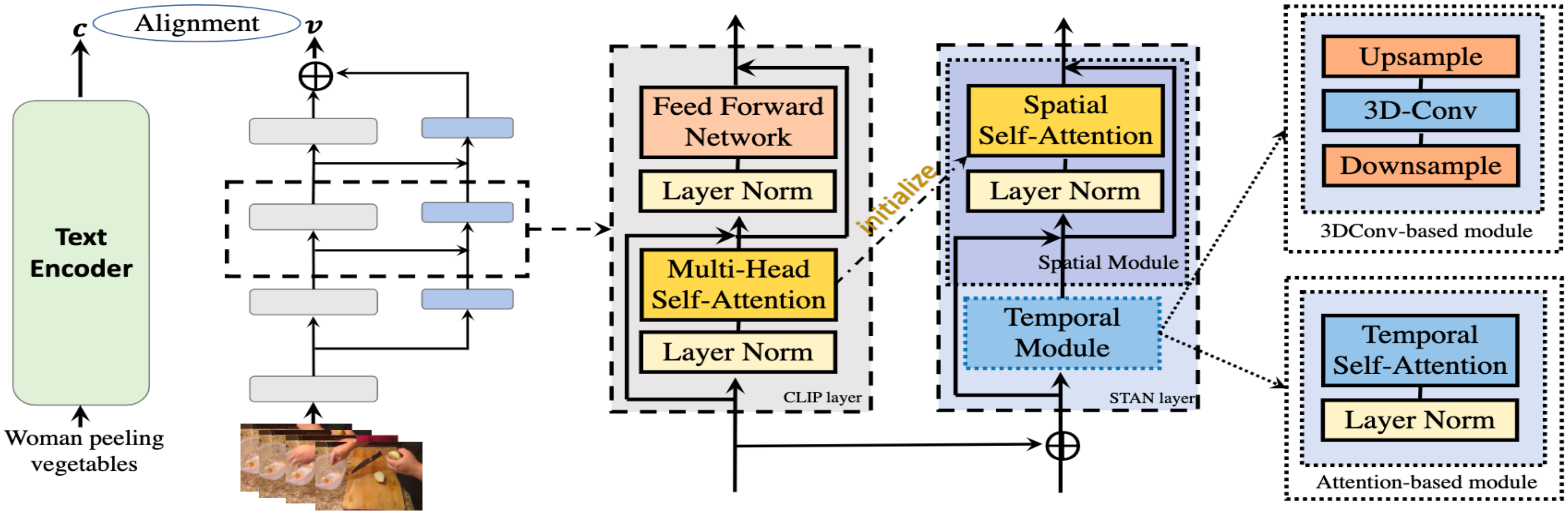
Key point for extending CLIP to the video domian?

temporal modeling + high-level knowledge + low-level knowledge

Branch structure：

- operate temporal modeling at different level of CLIP outputs
- the structure is outside the visual backbone avoiding to break the inherent structure of CLIP backbone and affect the pretrained high-level knowledge

Model design



Spatial-Temporal Auxiliary Network (STAN)

Branch structure with multi-level CLIP：

- attend to both high-level and low-level video representations

Separated spatial-temporal module：

- Computation efficiency
- Spatial module can reuse parameter in CLIP
- Present two instantiations of temporal module: Conv-based & Self-attention based

Experimental analysis

Table 2. Comparisons on MSR-VTT [43]. We train on Training-9K and test on Test-1k-A. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	22.0	46.8	59.9	6.0
Frozen [3]	31.0	59.5	70.5	3.0
HD-VILA [44]	35.6	65.3	78.0	3.0
All-in-one [40]	37.9	68.1	77.1	-
BridgeFormer [15]	37.6	64.8	75.1	3.0
Clover [18]	38.6	67.4	76.4	2.0
CLIP pretrained				
Clip4Clip [25]	44.5	71.4	81.6	2.0
CenterCLIP [47]	44.2	71.6	82.1	2.0
CLIP2Video* [13]	47.2	73.0	83.0	-
CAMoE* [9]	47.3	74.2	84.5	3.0
CLIP2TV-B/16 [14]	49.3	74.7	83.6	2.0
DRL-B/16* [42]	53.3	80.3	87.6	1.0
Our method				
A-STA-B/32	46.9	72.8	82.8	2.0
C-STA-B/32	46.6	72.8	82.2	2.0
A-STA-B/16	50.0	75.2	84.1	1.5
A-STA-B/16*	55.1	77.8	86.1	1.0

Table 5. Comparison on LSMDC [34]. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
Frozen [3]	15.0	30.8	40.3	20.0
HD-VILA [44]	17.4	34.1	44.1	15.0
BridgeFormer [15]	17.9	35.4	44.5	15.0
Clover [18]	22.7	42.0	52.6	9.0
CLIP pretrained				
Clip4Clip [25]	21.6	41.8	49.8	8.0
CAMoE* [9]	25.9	46.1	53.7	-
DRL-B/16 [42]	26.5	47.6	56.8	7.0
Our method				
A-STA-B/32	23.7	42.7	51.8	9.0
C-STA-B/32	23.1	42.2	51.0	9.0
A-STA-B/16	27.1	49.3	58.7	6.0
A-STA-B/16*	29.2	49.5	58.8	6.0

Table 4. Comparisons on DiDemo [1]. We concatenate all captions of a video into a single query. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	20.4	48.0	60.8	6.0
Frozen [3]	31.0	59.8	72.4	3.0
HD-VILA [44]	28.8	57.4	69.1	4.0
All-in-one [40]	32.7	61.4	73.5	3.0
BridgeFormer [15]	37.0	62.2	73.9	3.0
Clover [18]	48.6	74.3	82.2	2.0
CLIP pretrained				
Clip4Clip [25]	43.4	70.2	80.6	2.0
CAMoE* [9]	43.8	71.4	-	-
CLIP2TV [14]	45.5	69.7	80.6	2.0
DRL-B/16 [42]	49.0	76.5	84.5	2.0
Our method				
A-STA-B/32	46.2	70.4	80.0	2.0
C-STA-B/32	46.5	71.5	80.9	2.0
C-STA-B/16	49.4	74.9	83.2	1.0
C-STA-B/16*	54.6	78.4	85.1	1.0

Benchmark:

- MSRVTT
- DiDeMo
- LSMDC

Set-up:

- contrastive loss
- different model scale (B/16, B/32)
- w/ or w/o DSL

Experimental analysis

ByteDance 字节跳动

Table 2. Comparisons on MSR-VTT [43]. We train on Training-9K and test on Test-1k-A. * means extra tricks (*e.g.,* DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	22.0	46.8	59.9	6.0
Frozen [3]	31.0	59.5	70.5	3.0
HD-VILA [44]	35.6	65.3	78.0	3.0
All-in-one [40]	37.9	68.1	77.1	-
BridgeFormer [15]	37.6	64.8	75.1	3.0
Clover [18]	38.6	67.4	76.4	2.0
CLIP pretrained				
Clip4Clip [25]	44.5	71.4	81.6	2.0
CenterCLIP [47]	44.2	71.6	82.1	2.0
CLIP2Video* [13]	47.2	73.0	83.0	-
CAMoE* [9]	47.3	74.2	84.5	3.0
CLIP2TV-B/16 [14]	49.3	74.7	83.6	2.0
DRL-B/16* [42]	53.3	80.3	87.6	1.0
Our method				
A-STA-B/32	46.9	72.8	82.8	2.0
C-STA-B/32	46.6	72.8	82.2	2.0
A-STA-B/16	50.0	75.2	84.1	1.5
A-STA-B/16*	55.1	77.8	86.1	1.0

Table 5. Comparison on LSMDC [34]. * means extra tricks (*e.g.,* DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
Frozen [3]	15.0	30.8	40.3	20.0
HD-VILA [44]	17.4	34.1	44.1	15.0
BridgeFormer [15]	17.9	35.4	44.5	15.0
Clover [18]	22.7	42.0	52.6	9.0
CLIP pretrained				
Clip4Clip [25]	21.6	41.8	49.8	8.0
CAMoE* [9]	25.9	46.1	53.7	-
DRL-B/16 [42]	26.5	47.6	56.8	7.0
Our method				
A-STA-B/32	23.7	42.7	51.8	9.0
C-STA-B/32	23.1	42.2	51.0	9.0
A-STA-B/16	27.1	49.3	58.7	6.0
A-STA-B/16*	29.2	49.5	58.8	6.0

Table 4. Comparisons on DiDemo [1]. We concatenate all captions of a video into a single query. * means extra tricks (*e.g.,* DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	20.4	48.0	60.8	6.0
Frozen [3]	31.0	59.8	72.4	3.0
HD-VILA [44]	28.8	57.4	69.1	4.0
All-in-one [40]	32.7	61.4	73.5	3.0
BridgeFormer [15]	37.0	62.2	73.9	3.0
Clover [18]	48.6	74.3	82.2	2.0
CLIP pretrained				
Clip4Clip [25]	43.4	70.2	80.6	2.0
CAMoE* [9]	43.8	71.4	-	-
CLIP2TV [14]	45.5	69.7	80.6	2.0
DRL-B/16 [42]	49.0	76.5	84.5	2.0
Our method				
A-STA-B/32	46.2	70.4	80.0	2.0
C-STA-B/32	46.5	71.5	80.9	2.0
C-STA-B/16	49.4	74.9	83.2	1.0
C-STA-B/16*	54.6	78.4	85.1	1.0

Benchmark:
- MSRVTT
- DiDeMo
- LSMDC

Set-up:
- contrastive loss
- different model scale (B/16, B/32)
- w/ or w/o DSL

- SOTA on 3 video-text retrieval datasets under:
  - both model size B/32 and B/16
  - w/ and w/o extra tricks (e.g., DSL and QB-Norm)
  - outperform strong competitor including DRL, CAMoE, CenterCLIP

Table 2. Comparisons on MSR-VTT [43]. We train on Training-9K and test on Test-1k-A. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1↑	R@5↑	R@10↑	MdR↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	22.0	46.8	59.9	6.0
Frozen [3]	31.0	59.5	70.5	3.0
HD-VILA [44]	35.6	65.3	78.0	3.0
All-in-one [40]	37.9	68.1	77.1	-
BridgeFormer [15]	37.6	64.8	75.1	3.0
Clover [18]	38.6	67.4	76.4	2.0
CLIP pretrained				
Clip4Clip [25]	44.5	71.4	81.6	2.0
CenterCLIP [47]	44.2	71.6	82.1	2.0
CLIP2Video* [13]	47.2	73.0	83.0	-
CAMoE* [9]	47.3	74.2	84.5	3.0
CLIP2TV-B/16 [14]	49.3	74.7	83.6	2.0
DRL-B/16* [42]	53.3	80.3	87.6	1.0
Our method				
A-STA-B/32	46.9	72.8	82.8	2.0
C-STA-B/32	46.6	72.8	82.2	2.0
A-STA-B/16	50.0	75.2	84.1	1.5
A-STA-B/16*	55.1	77.8	86.1	1.0

Table 4. Comparisons on DiDemo [1]. We concatenate all captions of a video into a single query. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1↑	R@5↑	R@10↑	MdR↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	20.4	48.0	60.8	6.0
Frozen [3]	31.0	59.8	72.4	3.0
HD-VILA [44]	28.8	57.4	69.1	4.0
All-in-one [40]	32.7	61.4	73.5	3.0
BridgeFormer [15]	37.0	62.2	73.9	3.0
Clover [18]	48.6	74.3	82.2	2.0
CLIP pretrained				
Clip4Clip [25]	43.4	70.2	80.6	2.0
CAMoE* [9]	43.8	71.4	-	-
CLIP2TV [14]	45.5	69.7	80.6	2.0
DRL-B/16 [42]	49.0	76.5	84.5	2.0
Our method				
A-STA-B/32	46.2	70.4	80.0	2.0
C-STA-B/32	46.5	71.5	80.9	2.0
C-STA-B/16	49.4	74.9	83.2	1.0
C-STA-B/16*	54.6	78.4	85.1	1.0

ByteDance 字节跳动

Benchmark:

- MSRVTT
- DiDeMo
- LSMDC

Set-up:

- contrastive loss
- different model scale (B/16, B/32)
- w/ or w/o DSL

Table 5. Comparison on LSMDC [34]. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1↑	R@5↑	R@10↑	MdR↓
Pretrained on large-scale video-text dataset				
Frozen [3]	15.0	30.8	40.3	20.0
HD-VILA [44]	17.4	34.1	44.1	15.0
BridgeFormer [15]	17.9	35.4	44.5	15.0
Clover [18]	22.7	42.0	52.6	9.0
CLIP pretrained				
Clip4Clip [25]	21.6	41.8	49.8	8.0
CAMoE* [9]	25.9	46.1	53.7	-
DRL-B/16 [42]	26.5	47.6	56.8	7.0
Our method				
A-STA-B/32	23.7	42.7	51.8	9.0
C-STA-B/32	23.1	42.2	51.0	9.0
A-STA-B/16	27.1	49.3	58.7	6.0
A-STA-B/16*	29.2	49.5	58.8	6.0

- Obvious advantage over posterior structure method with comparable model size i.e., CLIP4clip (+2.9% averaged on 3 datasets)

Table 2. Comparisons on MSR-VTT [43]. We train on Training-9K and test on Test-1k-A. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	22.0	46.8	59.9	6.0
Frozen [3]	31.0	59.5	70.5	3.0
HD-VILA [44]	35.6	65.3	78.0	3.0
All-in-one [40]	37.9	68.1	77.1	-
BridgeFormer [15]	37.6	64.8	75.1	3.0
Clover [18]	38.6	67.4	76.4	2.0
CLIP pretrained				
Clip4Clip [25]	44.5	71.4	81.6	2.0
CenterCLIP [47]	44.2	71.6	82.1	2.0
CLIP2Video* [13]	47.2	73.0	83.0	-
CAMoE* [9]	47.3	74.2	84.5	3.0
CLIP2TV-B/16 [14]	49.3	74.7	83.6	2.0
DRL-B/16* [42]	53.3	80.3	87.6	1.0
Our method				
A-STA-B/32	46.9	72.8	82.8	2.0
C-STA-B/32	46.6	72.8	82.2	2.0
A-STA-B/16	50.0	75.2	84.1	1.5
A-STA-B/16*	55.1	77.8	86.1	1.0

Table 4. Comparisons on DiDemo [1]. We concatenate all captions of a video into a single query. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	20.4	48.0	60.8	6.0
Frozen [3]	31.0	59.8	72.4	3.0
HD-VILA [44]	28.8	57.4	69.1	4.0
All-in-one [40]	32.7	61.4	73.5	3.0
BridgeFormer [15]	37.0	62.2	73.9	3.0
Clover [18]	48.6	74.3	82.2	2.0
CLIP pretrained				
Clip4Clip [25]	43.4	70.2	80.6	2.0
CAMoE* [9]	43.8	71.4	-	-
CLIP2TV [14]	45.5	69.7	80.6	2.0
DRL-B/16 [42]	49.0	76.5	84.5	2.0
Our method				
A-STA-B/32	46.2	70.4	80.0	2.0
C-STA-B/32	46.5	71.5	80.9	2.0
C-STA-B/16	49.4	74.9	83.2	1.0
C-STA-B/16*	54.6	78.4	85.1	1.0

Benchmark:

- MSRVTT
- DiDeMo
- LSMDC

Set-up:

- contrastive loss
- different model scale (B/16, B/32)
- w/ or w/o DSL

Table 5. Comparison on LSMDC [34]. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
Frozen [3]	15.0	30.8	40.3	20.0
HD-VILA [44]	17.4	34.1	44.1	15.0
BridgeFormer [15]	17.9	35.4	44.5	15.0
Clover [18]	22.7	42.0	52.6	9.0
CLIP pretrained				
Clip4Clip [25]	21.6	41.8	49.8	8.0
CAMoE* [9]	25.9	46.1	53.7	-
DRL-B/16 [42]	26.5	47.6	56.8	7.0
Our method				
A-STA-B/32	23.7	42.7	51.8	9.0
C-STA-B/32	23.1	42.2	51.0	9.0
A-STA-B/16	27.1	49.3	58.7	6.0
A-STA-B/16*	29.2	49.5	58.8	6.0

In STAN, 3D convolution based temporal module is comparable (+-0.3) with self-attention based temporal module when transferring to smaller datasets, e.g., MSRVTT and DiDeMo; self-attention based temporal module is better for larger scale dataset, e,g., LSMDC (+0.6)

Table 2. Comparisons on MSR-VTT [43]. We train on Training-9K and test on Test-1k-A. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	22.0	46.8	59.9	6.0
Frozen [3]	31.0	59.5	70.5	3.0
HD-VILA [44]	35.6	65.3	78.0	3.0
All-in-one [40]	37.9	68.1	77.1	-
BridgeFormer [15]	37.6	64.8	75.1	3.0
Clover [18]	38.6	67.4	76.4	2.0
CLIP pretrained				
Clip4Clip [25]	44.5	71.4	81.6	2.0
CenterCLIP [47]	44.2	71.6	82.1	2.0
CLIP2Video* [13]	47.2	73.0	83.0	-
CAMoE* [9]	47.3	74.2	84.5	3.0
CLIP2TV-B/16 [14]	49.3	74.7	83.6	2.0
DRL-B/16* [42]	53.3	80.3	87.6	1.0
Our method				
A-STA-B/32	46.9	72.8	82.8	2.0
C-STA-B/32	46.6	72.8	82.2	2.0
A-STA-B/16	50.0	75.2	84.1	1.5
A-STA-B/16*	55.1	77.8	86.1	1.0

Table 4. Comparisons on DiDemo [1]. We concatenate all captions of a video into a single query. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
ClipBERT [22]	20.4	48.0	60.8	6.0
Frozen [3]	31.0	59.8	72.4	3.0
HD-VILA [44]	28.8	57.4	69.1	4.0
All-in-one [40]	32.7	61.4	73.5	3.0
BridgeFormer [15]	37.0	62.2	73.9	3.0
Clover [18]	48.6	74.3	82.2	2.0
CLIP pretrained				
Clip4Clip [25]	43.4	70.2	80.6	2.0
CAMoE* [9]	43.8	71.4	-	-
CLIP2TV [14]	45.5	69.7	80.6	2.0
DRL-B/16 [42]	49.0	76.5	84.5	2.0
Our method				
A-STA-B/32	46.2	70.4	80.0	2.0
C-STA-B/32	46.5	71.5	80.9	2.0
C-STA-B/16	49.4	74.9	83.2	1.0
C-STA-B/16*	54.6	78.4	85.1	1.0

ByteDance 字节跳动

Benchmark:
- MSRVTT
- DiDeMo
- LSMDC

Set-up:
- contrastive loss
- different model scale (B/16, B/32)
- w/ or w/o DSL

Table 5. Comparison on LSMDC [34]. * means extra tricks (*e.g.*, DSL [9] and QB-Norm [5]) are utilized during inference.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Pretrained on large-scale video-text dataset				
Frozen [3]	15.0	30.8	40.3	20.0
HD-VILA [44]	17.4	34.1	44.1	15.0
BridgeFormer [15]	17.9	35.4	44.5	15.0
Clover [18]	22.7	42.0	52.6	9.0
CLIP pretrained				
Clip4Clip [25]	21.6	41.8	49.8	8.0
CAMoE* [9]	25.9	46.1	53.7	-
DRL-B/16 [42]	26.5	47.6	56.8	7.0
Our method				
A-STA-B/32	23.7	42.7	51.8	9.0
C-STA-B/32	23.1	42.2	51.0	9.0
A-STA-B/16	27.1	49.3	58.7	6.0
A-STA-B/16*	29.2	49.5	58.8	6.0

- Simple and potentially compatiable with other SOTA methods (future work)
  - multi-modal interaction modeling
  - hard sample matching
  - hierarchical modeling

Experimental analysis

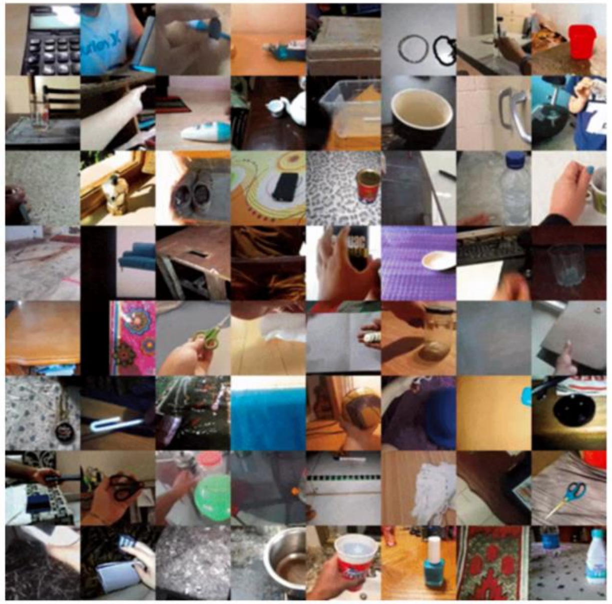
Table 6. Comparison between our method and the state-of-the-arts on Kinetics-400 validation set [21]. We report the FLOPs of all views.

Methods	Pretrain	Frames	Testing Views	GFLOPs	Top-1 Accuracy	Top-5 Accuracy
Large-scale image pretraining						
VTN-ViT-B [28]	ImageNet-21 K	250	1 × 1	3992	78.6	93.7
TimeSformer-L [4]	ImageNet-21 K	96	1 × 3	7140	80.7	94.7
Mformer-HR [32]	ImageNet-21 K	16	10 × 3	28764	83.1	95.9
Swin-L (384 ↑)1001[24]	ImageNet-21 K	32	10 × 5	105350	84.9	96.7
MViTv2-L (312 ↑)1001[23]	ImageNet-21 K	40	5 × 3	42420	**86.1**	97.0
ViViT-H [2]	JFT-300M	32	4 × 3	17352	84.8	95.8
TokenLearner-L/10 [35]	JFT-300M	-	4 × 3	48912	85.4	96.3
Large-scale image-text pretraining						
CLIP-B/16 [11]	CLIP-400M	8	4 × 3	-	81.1	94.8
Action-CLIP-B/16 [41]	CLIP-400M	32	10 × 3	16890	83.8	96.2
A6 [20]	CLIP-400M	16	–	-	76.9	93.5
STadapter-CLIP-B/16 [41]	CLIP-400M	8	1 × 3	455	82.0	95.7
STadapter-CLIP-B/16 [41]	CLIP-400M	32	1 × 3	1821	82.7	96.2
X-CLIP-B/16 [41]	CLIP-400M	8	4 × 3	1740	83.8	96.7
X-CLIP-B/16 [41]	CLIP-400M	16	4 × 3	3444	84.7	96.8
Our method						
C-STA-B/16	CLIP-400M	8	1 × 3	714	83.1	96.0
A-STA-B/16	CLIP-400M	8	1 × 3	593	84.2	96.5
A-STA-B/16	CLIP-400M	16	1 × 3	1187	**84.9**	96.7

Kinetics-400:

- Be superior to CLIP-based method on both acc and FLOPs
- Comparable acc with much lower FLOPs than Image-pretrained SOTA.

Experimental analysis



Sample classes in Something-Something-v2

- Putting something on a surface
- Moving something up
- Covering something with something
- Pushing something from left to right
- Moving something down
- Pushing something from right to left
- Uncovering something

Table 5. Comparison on Something-Something-v2 validation set [15]. We report the FLOPs of all views. * means our implementation.

Methods	Pretrain	Frames	Testing Views	GFLOPs	Top-1 Accuracy	Top-5 Accuracy
TimeSformer-HR [4]	ImageNet-21 K	16	1 × 3	5109	62.5	-
ViViT-L [2]	K400	16	4 × 3	11892	65.4	89.8
MViT-B-24 [21]	K600	32	1 × 3	708	68.7	91.5
Video-Swin-B [24]	K400	32	1 × 3	963	**69.6**	**92.7**
CLIP-B/16 [9]	CLIP-400M	8	1 × 3	-	44.0	76.2
X-CLIP-B/16* [40]	CLIP-400M	8	1 × 3	435	63.1	89.0
STadapter-CLIP-B/16 [40]	CLIP-400M	8	1 × 3	489	67.1	91.2
STadapter-CLIP-B/16 [40]	CLIP-400M	32	1 × 3	1955	69.5	92.6
Our method						
STAN-conv-B/16	CLIP-400M	8	1 × 3	845	65.2	90.5
STAN-self-B/16	CLIP-400M	8	1 × 3	688	67.6	91.4
STAN-self-B/16	CLIP-400M	16	1 × 3	1376	69.5	**92.7**

SSv2:

- CLIP has little advantage
- significantly improve CLIP-baseline (+21%); best among CLIP-based method
- Comparable with video-pretrained method.

Experimental analysis

Table 1. Ablation studies on different datasets. For MSRVTT and DiDemo, we use CLIP-B/32 as the baseline and report Recall@1; for K400 and SSv2, we use CLIP-B/16 as the baseline and report Top1 Accuracy. We adopt temporal attention as our Cross-Frame module.

Components				Results			
Cross-Frame	Intra-Frame	Branch structure	Multi-level	MSR-VTT	DiDemo	K400	SSv2
				43.1	43.4	79.9	44
✓	✓			43.9	43.5	80.3	55.9
✓	✓	✓		44.2	43.6	80.8	58.6
	✓	✓	✓	44.3	44.5	81.0	48.1
✓		✓	✓	43.1	43.7	80.0	55.7
✓	✓	✓	✓	46.9	46.2	82.6	65.9
+ Testing Techniques (DSL [9] or 1 × 3-views)				**49.7**	**51.4**	84.2	67.6

Ablation on different components

- remove branch structure and multi-level, performance dicreased a lot - > branch structure and multi-level is important
- removing spatial module has more impact on MSRVTT/Didemo/K400 than Sthsthv2 -> spatial modeling is benefitial to low-level knowledge transfer, low-level knowledge is benfitial to both recogntion and retrieval
- temporal module >+10% on sthsthv2, s-t module +branch structure+multi-level >+20 -> temporal module is effective, branch structure and multi-level imporve temporal modeling

Thanks!

https://arxiv.org/abs/2301.11116
https://github.com/farewellthree/STAN