The 18th European Conference on Computer Vision ECCV 2024

Leveraging Temporal Contextualization for Video Action Recognition

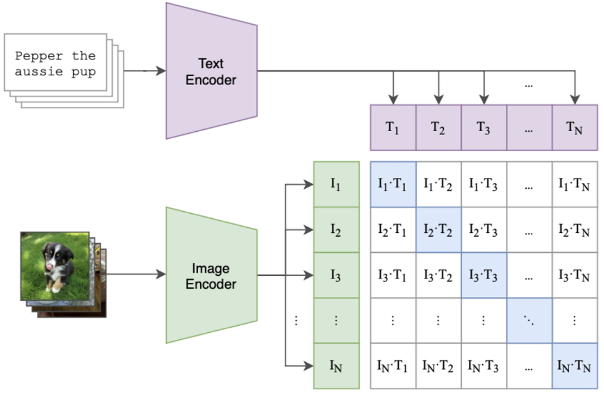
Minji Kim[1†] Dongyoon Han[2] Taekyung Kim[2*] Bohyung Han[1*]

[1] Seoul National University
[2] NAVER AI Lab

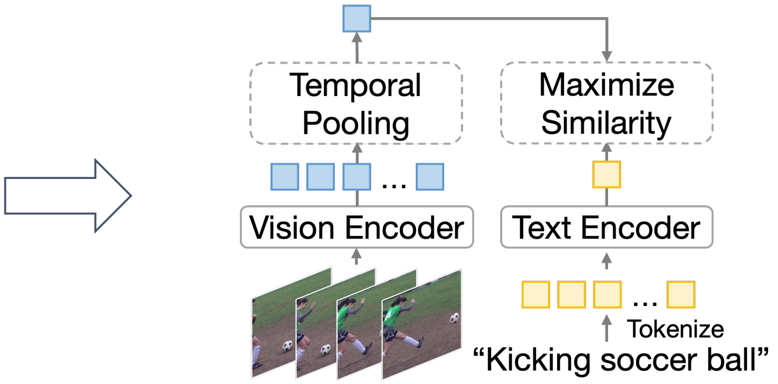
† Work done during an internship at NAVER AI Lab
* Corresponding authors

Background

- **Fine-tuning image-based VLMs (e.g., CLIP) for video action recognition** enables open-vocabulary generalization w/o expensive video-text pretraining
- A naïve baseline: **frame-wise attention**
  → Limitation: **no token interactions** in the **temporal** axis



Contrastive Language-Image Pretraining (CLIP)



Fine-tune CLIP with video-text pairs

Background

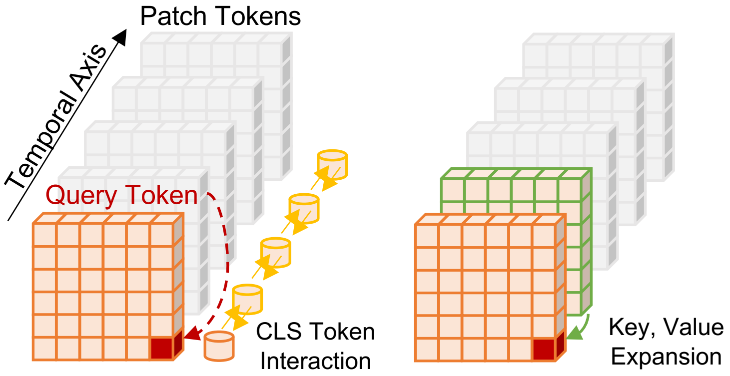
- To consider **temporal cues** during the frame-wise representation encoding, previous works additionally incorporate **reference tokens:**

$$\underbrace{\mathbf{z}_t^l = f_{\theta_v}^l(\mathbf{z}_t^{l-1}}_{\text{t-th frame patch tokens}}, \underbrace{\mathbf{s}^{l-1}}_{\text{reference tokens}})$$

- However, these reference tokens are *insufficient* for proper temporal modeling

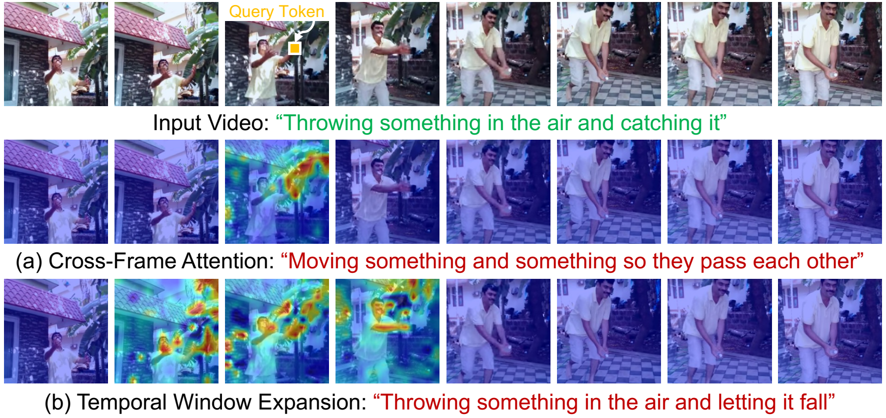
Limitation of Previous Temporal Modeling

- **Short-range token interactions** *hinder* models capturing essential temporal dynamics
- → <u>We need global interactions</u> to achieve better video representations!



(a) Cross-Frame Attention

(b) Temporal Window Expansion

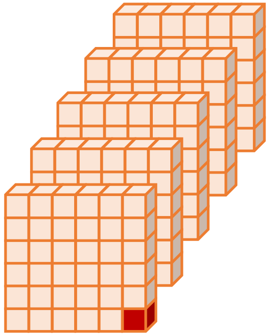


Reference Tokens	CLS tokens from all frames	Patch tokens from adjacent frames

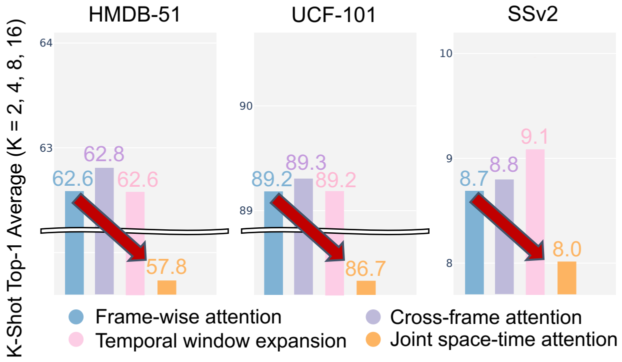
→ **Short-range token interactions**
Fail to capture essential video information

Limitation of Previous Temporal Modeling

- A **naïve** approach for global interactions: using ***all*** patch tokens as a reference
- Problem: extending CLIP's temporal sequence length **degrades attention quality** because it wasn't trained on long sequences



(c) Joint Space-Time Attention



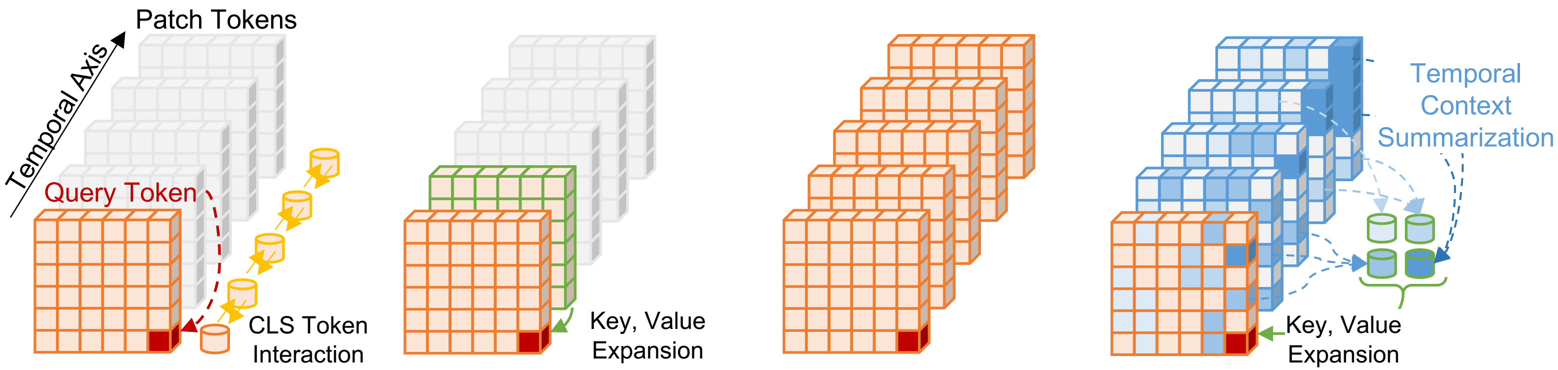
Patch tokens from all frames

⇨

😵 **Extrapolation** challenge
Costly / Suboptimal performance

Solution: Temporal Contextualization

- Key Idea: **Summarize informative tokens from the entire video** into a small set of tokens, called ***context tokens***, and **reference** them during feature encoding

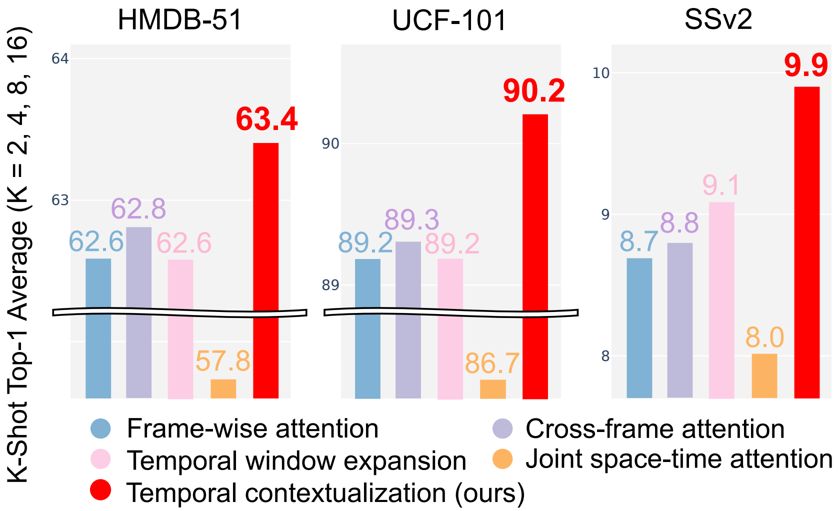


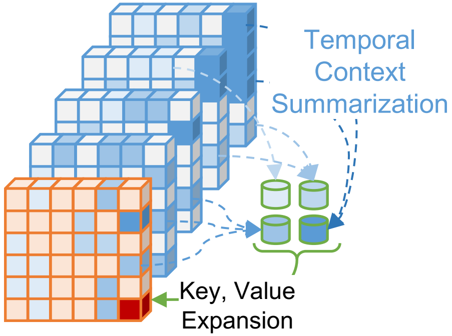
(a) Cross-Frame Attention
(b) Temporal Window Expansion
(c) Joint Space-Time Attention
(d) Temporal Contextualization (Ours)

Reference Tokens	CLS tokens from all frames	Patch tokens from adjacent frames	Patch tokens from all frames	**Context tokens**
	Short-range token interactions Fail to capture essential video information		**Extrapolation** challenge Costly / Suboptimal	Deliver **global** information Maintain CLIP's **effective length**

Solution: Temporal Contextualization

- Key Idea: **Summarize informative tokens from the entire video** into a small set of tokens, called ***context tokens***, and **reference** them during feature encoding





(d) Temporal Contextualization (Ours)

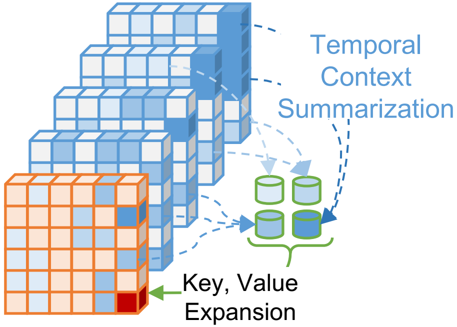
Context tokens

Deliver **global** information
Maintain CLIP's **effective length**

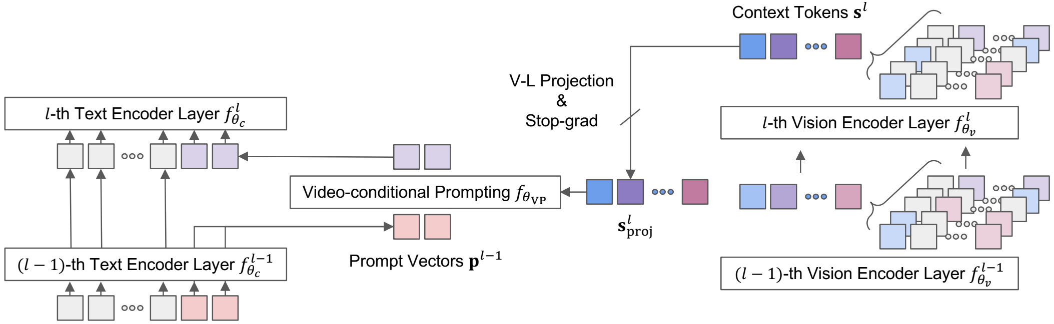
Using ***context tokens*** as a reference during the feature encoding **consistently improves** action recognition performance.

Temporally Contextualized CLIP (TC-CLIP)

- A novel paradigm of **extending CLIP to videos** by encoding ***holistic*** video information through advanced temporal analysis
  1. **Temporal Contextualization (TC):** allows ***global interactions*** by **summarizing** pivotal video information into **context tokens** and **referencing** them during the encoding process
  2. **Video-conditional Prompting (VP):** injects ***instance contexts*** into text modality to support **lack of textual semantics** in action recognition benchmarks
  3. **Solid performance**: TC-CLIP achieves **SOTA** on diverse benchmarks & protocols



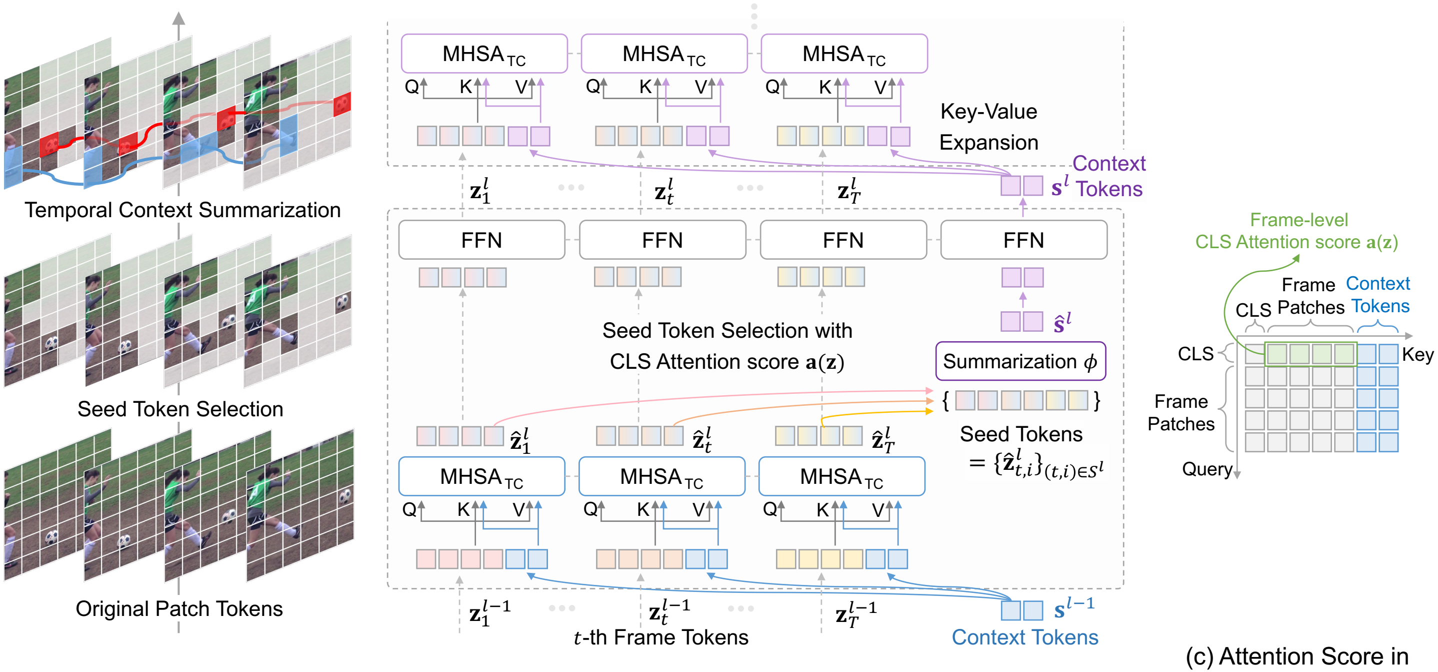
Temporal Contextualization (TC)



Video-conditional Prompting (VP)

Temporal Contextualization (TC)

- A **layer-wise temporal information infusion** mechanism for videos
- Three steps of TC



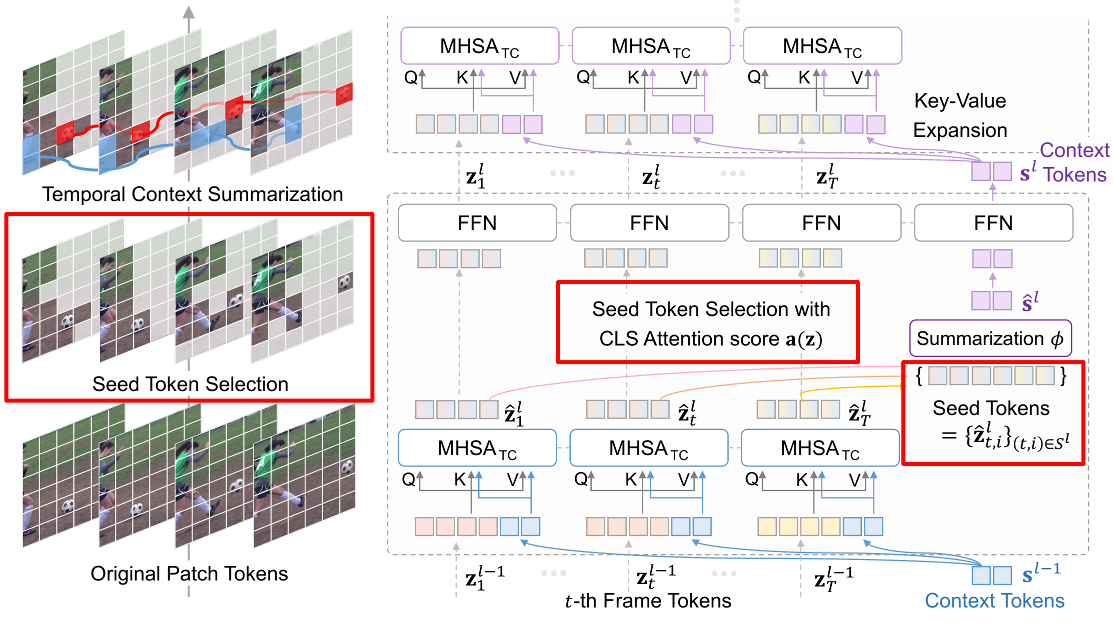
(a) Overall TC Pipeline

(b) Encoding Process of TC

(c) Attention Score in MHSA_{TC}

Temporal Contextualization (TC)

- Step 1) **Informative token selection** in each frame
  - To avoid ***redundant*** tokens in videos, we select **seed tokens** by using **CLS attention scores** obtained from self-attention operation **in each frame** as criteria



(a) Overall TC Pipeline

(b) Encoding Process of TC

query of CLS token

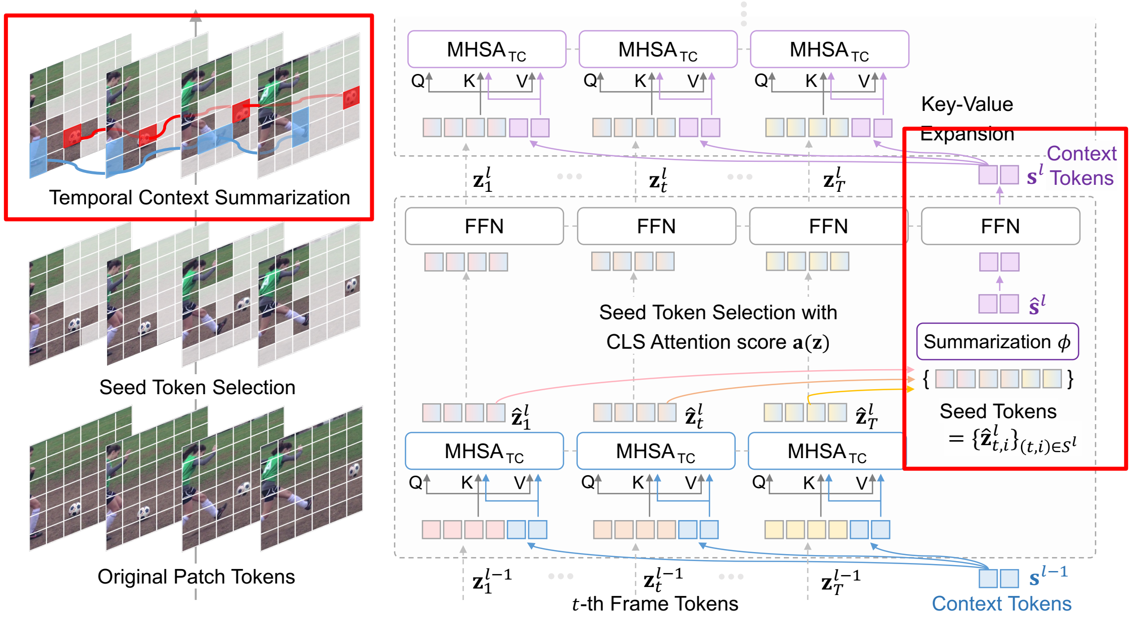
keys of patch tokens

$$\mathbf{a}(\mathbf{z}_t) = \mathrm{Softmax}\left(\frac{\mathbf{q}_{\mathrm{cls}}\mathbf{K}_{\mathbf{z}_t}^{\mathrm{T}}}{\sqrt{d}}\right)$$

t-th frame patch tokens

Temporal Contextualization (TC)

- Step 2) Spatio-temporal context **summarization**
  - To obtain **context tokens**, **cluster and merge** all the seed tokens from all frames by using token aggregation function



(a) Overall TC Pipeline

(b) Encoding Process of TC

aggregation function

patch token after self-attention

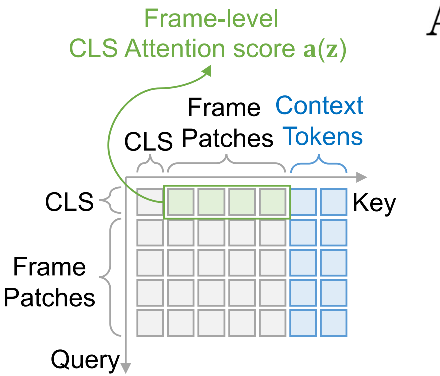
$$\hat{\mathbf{s}} = \phi\left(\{\hat{\mathbf{z}}_{t,i}\}_{(t,i)\in\mathcal{S}}\right)$$

$$\mathcal{S} = \{(t,i) | i \in \mathcal{S}_t, t = 1, \ldots, T\}$$

seed token indices from all frames

Temporal Contextualization (TC)

- Step 3) **Temporal context infusion**
  - Finally, the summarized context is **infused** to all patch tokens by **expanding key-value pairs**:



Attention Score in $\mathrm{MHSA}_{\mathrm{TC}}$

KV of context tokens

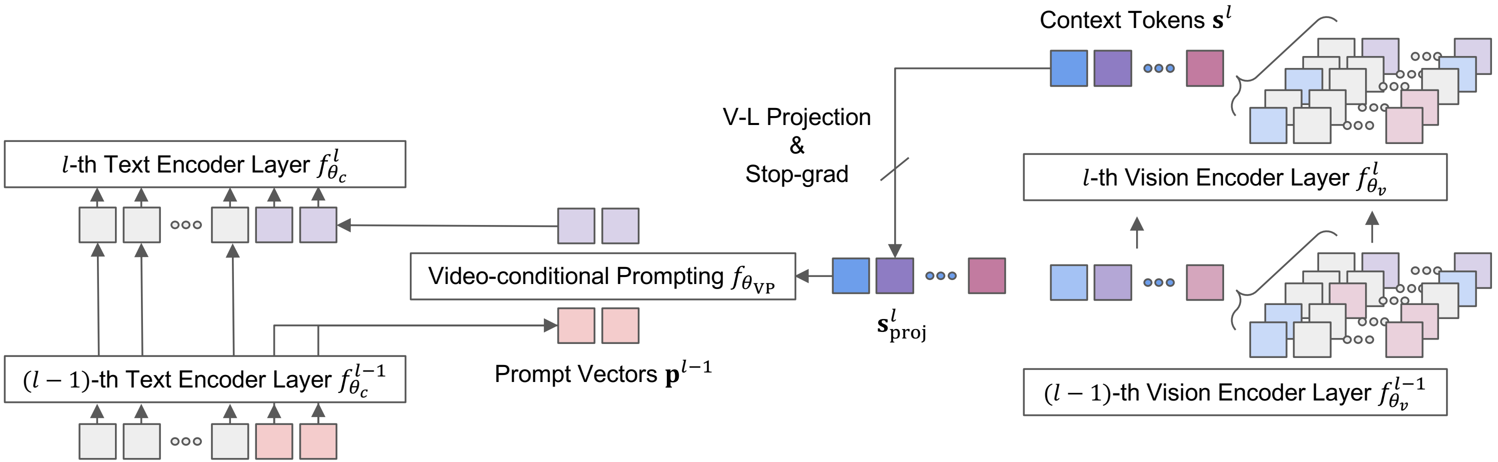
$$\mathrm{Attention}_{\mathrm{TC}}(\mathbf{z}_t, \mathbf{s}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{\mathbf{z}_t}\left[\mathbf{K}_{\mathbf{z}_t} | \mathbf{K}_{\mathbf{s}}\right]^{\mathsf{T}}}{\sqrt{d}} + \mathbf{B}\right)\left[\mathbf{V}_{\mathbf{z}_t} | \mathbf{V}_{\mathbf{s}}\right]$$

QKV of t-th frame patch tokens

Learnable bias $\mathbf{B}_{ij} = \begin{cases} b_{\mathrm{local}} & \text{if } j \leq N+1 \\ b_{\mathrm{global}} & \text{otherwise,} \end{cases}$

Video-conditional Prompting (VP)

- Generates **instance-level textual prompts** that support the **lack of textual semantics** in action recognition datasets, where category names are the only description of actions *(e.g., skateboarding, skydiving, ski jumping)*
- **Video information from the context tokens** is **injected** to the **text prompt** vectors based on a cross-attention mechanism



Video-conditional Prompting (VP)

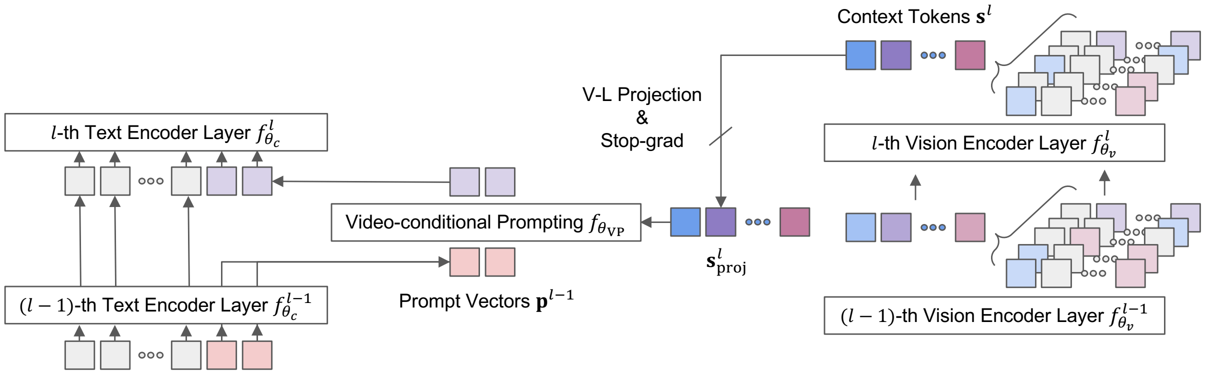
- We perform video-conditional prompting before the last text encoder layer:

prompt vectors

class name tokens

$$[\mathbf{p}^l, \mathbf{c}^l] = \begin{cases} f^l_{\theta_c}([f_{\theta_{\mathrm{VP}}}(\mathbf{p}^{l-1}, \mathbf{s}^l_{\mathrm{proj}}), \mathbf{c}^{l-1}]) & \text{if } l = L_c \\ f^l_{\theta_c}([\mathbf{p}^{l-1}, \mathbf{c}^{l-1}]) & \text{otherwise.} \end{cases}$$

$$\hat{\mathbf{p}}^{l-1} = \mathrm{MHCA}(\mathrm{LN}_p(\mathbf{p}^{l-1}), \mathrm{LN}_s(\mathbf{s}^l_{\mathrm{proj}})) + \mathbf{p}^{l-1}$$
$$\tilde{\mathbf{p}}^{l-1} = \mathrm{FFN}(\mathrm{LN}(\hat{\mathbf{p}}^{l-1})) + \hat{\mathbf{p}}^{l-1}$$



Experiments

- **SOTA** performance on **zero/few-shot, base-to-novel, fully-supervised** action recognition

Zero-shot action recognition

Method	WE	HMDB-51	UCF-101	K600 (Top-1)	K600 (Top-5)	All (Top-1)
Vanilla CLIP [32]		40.8 ± 0.3	63.2 ± 0.2	59.8 ± 0.3	83.5 ± 0.2	54.6
ActionCLIP [39]†		49.1 ± 0.4	68.0 ± 0.9	56.1 ± 0.9	83.2 ± 0.2	57.7
A5 [14]		44.3 ± 2.2	69.3 ± 4.2	55.8 ± 0.7	81.4 ± 0.3	56.5
X-CLIP [29]		44.6 ± 5.2	72.0 ± 2.3	65.2 ± 0.4	86.1 ± 0.8	60.6
Vita-CLIP [41]		48.6 ± 0.6	75.0 ± 0.6	67.4 ± 0.5	-	63.7
ViFi-CLIP [34]†		52.3 ± 0.2	78.9 ± 1.1	70.7 ± 0.8	92.1 ± 0.3	67.3
TC-CLIP (Ours)		**53.7** ± 0.7	**80.4** ± 0.9	**72.7** ± 0.5	**93.2** ± 0.2	**68.9**
ActionCLIP [39]†	✓	51.9 ± 0.5	74.2 ± 1.0	67.5 ± 1.2	90.7 ± 0.1	64.5
ViFi-CLIP [34]†	✓	52.2 ± 0.7	81.0 ± 0.9	73.9 ± 0.5	93.3 ± 0.3	69.0
Open-VCLIP [42]	✓	53.9 ± 1.2	**83.4** ± 1.2	73.0 ± 0.8	93.2 ± 0.1	70.1
TC-CLIP (Ours)	✓	**54.2** ± 0.7	82.9 ± 0.6	**75.8** ± 0.5	**94.4** ± 0.2	**71.0**
Using LLM-based text augmentation						
MAXI [24]	✓	52.3 ± 0.7	78.2 ± 0.8	71.5 ± 0.8	92.5 ± 0.4	67.3
OST [4]	✓	55.9 ± 1.2	79.7 ± 1.1	75.1 ± 0.6	94.6 ± 0.2	70.2
FROSTER [10]	✓	54.8 ± 1.3	84.8 ± 1.1	74.8 ± 0.9	-	71.5
TC-CLIP (Ours)	✓	**56.0** ± 0.3	**85.4** ± 0.8	**78.1** ± 1.0	**95.7** ± 0.3	**73.2**

Few-shot action recognition

Method	HMDB-51 K=2	K=4	K=8	K=16	UCF-101 K=2	K=4	K=8	K=16	SSv2 K=2	K=4	K=8	K=16	All Avg.
Vanilla CLIP [32]	41.9	41.9	41.9	41.9	63.6	63.6	63.6	63.6	2.7	2.7	2.7	2.7	36.1
ActionCLIP [39]	47.5	57.9	57.3	59.1	70.6	71.5	73.0	91.4	4.1	5.8	8.4	11.1	46.5
A5 [14]	39.7	50.7	56.0	62.4	71.4	79.9	85.7	89.9	4.4	5.1	6.1	9.7	46.8
X-CLIP [29]	53.0	57.3	62.8	64.0	76.4	83.4	88.3	91.4	3.9	4.5	6.8	10.0	50.2
ViFi-CLIP [34]	57.2	**62.7**	64.5	66.8	80.7	85.1	90.0	92.7	6.2	7.4	8.5	12.4	52.9
TC-CLIP (Ours)	**57.3**	62.3	**67.3**	**68.6**	**85.9**	**89.9**	**92.5**	**94.6**	**7.3**	**8.6**	**9.3**	**14.0**	**54.8**
Using LLM-based text augmentation													
OST [4]	**59.1**	62.9	64.9	68.2	82.5	87.5	91.7	93.9	7.0	7.7	8.9	12.2	53.9
TC-CLIP (Ours)	58.6	**63.3**	**65.5**	**68.8**	**86.8**	**90.1**	**92.0**	**94.3**	**7.3**	**8.6**	**9.3**	**14.0**	**54.9**

Base-to-novel generalization

Method	K-400 Base	Novel	HM	HMDB-51 Base	Novel	HM	UCF-101 Base	Novel	HM	SSv2 Base	Novel	HM	All (Avg.) Base	Novel	HM
Vanilla CLIP [32]	62.3	53.4	57.5	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1	49.8	42.3	45.7
ActionCLIP [39]	61.0	46.2	52.6	69.1	37.3	48.5	90.1	58.1	70.7	13.3	10.1	11.5	58.5	37.9	46.0
A5 [14]	69.7	37.6	48.8	46.2	16.0	23.8	90.5	40.4	55.8	8.3	5.3	6.4	53.7	24.8	33.9
X-CLIP [29]	74.1	56.4	64.0	69.4	45.5	55.0	89.9	58.9	71.2	8.5	6.6	7.4	60.5	41.9	49.5
ViFi-CLIP [34]	76.4	61.1	67.9	**73.8**	53.3	61.9	92.9	67.7	78.3	16.2	12.1	13.9	64.8	48.6	55.5
Open-VCLIP [42]†	76.5	62.6	68.9	70.3	50.4	58.9	94.8	77.5	85.3	16.0	11.0	13.0	64.4	50.4	56.5
TC-CLIP (Ours)	**78.9**	**63.6**	**70.4**	73.3	**54.1**	**62.2**	**95.5**	**78.0**	**85.9**	**17.5**	**13.4**	**15.2**	**66.3**	**52.3**	**58.5**
Using LLM-based text augmentation															
FROSTER [10]	77.8	64.3	70.4	**74.1**	58.0	65.1	95.3	80.0	87.0	**18.3**	12.2	14.6	**66.4**	53.6	59.3
TC-CLIP (Ours)	**79.1**	**65.4**	**71.6**	73.3	**59.1**	**65.5**	**95.4**	**81.6**	**88.0**	17.5	**13.4**	**15.2**	66.3	**54.9**	**60.1**

Fully-supervised action recognition

Method	Top-1	Top-5	F	Views
ActionCLIP [39]	83.8	96.2	32	10×3
X-CLIP [29]	84.7	96.8	16	4×3
Vita-CLIP [41]	82.9	96.3	16	4×3
ViFi-CLIP [34]	83.9	96.3	16	4×3
OST [4]	83.2	-	16	1×1
TC-CLIP (Ours)	**85.2**	**96.9**	16	4×3

Analysis

- Component-wise ablation: **TC** and **VP** are **both effective**

Case	Without weight-space ensembling: HMDB-51	UCF-101	K-600	All (Δ)	With weight-space ensembling: HMDB-51	UCF-101	K-600	All (Δ)
Baseline	52.3 ± 0.2	78.9 ± 1.1	70.7 ± 0.8	67.3	52.2 ± 0.7	81.0 ± 0.9	73.9 ± 0.5	69.0
(a) +TC	53.6 ± 0.2	78.6 ± 1.0	71.8 ± 0.7	68.0 (+0.7)	54.3 ± 0.6	81.9 ± 1.0	75.5 ± 1.0	70.6 (+1.6)
(b) +VP	53.2 ± 0.8	80.5 ± 0.7	71.6 ± 0.9	68.4 (+1.1)	53.4 ± 0.8	82.0 ± 0.9	74.7 ± 0.7	70.0 (+1.0)
(c) +TC+VP	53.7 ± 0.7	80.4 ± 0.9	72.7 ± 0.5	68.9 (+1.6)	54.2 ± 1.1	82.9 ± 0.9	75.8 ± 0.4	71.0 (+2.0)

- TC is **robust** across **diverse token aggregation strategies**

(a) Seed token selection strategy.

Case	HMDB	UCF	SSv2	All (Δ)
Baseline	62.6	89.2	8.7	53.5
No selection	62.8	89.8	9.7	54.1 (+0.6)
Head-wise key norm	62.3	89.8	9.8	54.0 (+0.5)
Averaged key norm	62.5	89.4	9.3	53.7 (+0.2)
Head-wise CLS attn.	63.4	89.9	9.7	54.3 (+0.8)
Averaged CLS attn.	63.4	90.2	**9.9**	**54.5** (+1.0)
Patch saliency [5]	62.9	**90.3**	9.6	54.2 (+0.7)
ATS [8]	**63.5**	**90.3**	9.8	**54.5** (+1.0)

(b) Context token summarization strategy.

Case	HMDB	UCF	SSv2	All (Δ)
Baseline	62.6	89.2	8.7	53.5
No merge	57.2	85.6	7.7	50.2 (−3.3)
Random merge	58.8	87.1	7.5	51.2 (−2.3)
K-means [25]	62.1	89.7	9.0	53.6 (+0.1)
DPC-KNN [13]	63.3	**90.2**	9.8	54.4 (+0.9)
Bipartite soft matching [1, 15]	**63.4**	**90.2**	**9.9**	54.5 (+1.0)
Bipartite w/ attention weights	62.9	89.8	**9.9**	54.2 (+0.7)
Bipartite w/ saliency weights [5]	62.4	89.9	9.6	54.0 (+0.5)

Analysis

- **Learnable bias** in MHSA_{TC} is helpful to distinguish local/global information
- TC is not sensitive to the choice of seed token ratio and the number of context tokens

(a) Positional embedding design.

Case	HMDB	UCF	SSv2	All
Spatial embedding	62.9	90.0	9.8	54.2
Joint space-time embedding	63.2	**90.2**	9.8	54.4
Spatial embedding + Bias	**63.4**	**90.2**	**9.9**	**54.5**
Joint embedding + Bias	62.9	**90.2**	9.8	54.3

(b) Seed token ratio α.

α	HMDB	UCF	SSv2	All
0.2	62.6	90.1	9.8	54.2
0.3	**63.4**	90.2	**9.9**	54.5
0.4	63.2	**90.4**	9.8	54.5
0.5	63.3	90.3	9.8	54.5
0.6	63.1	90.2	9.8	54.4

(c) Context token k.

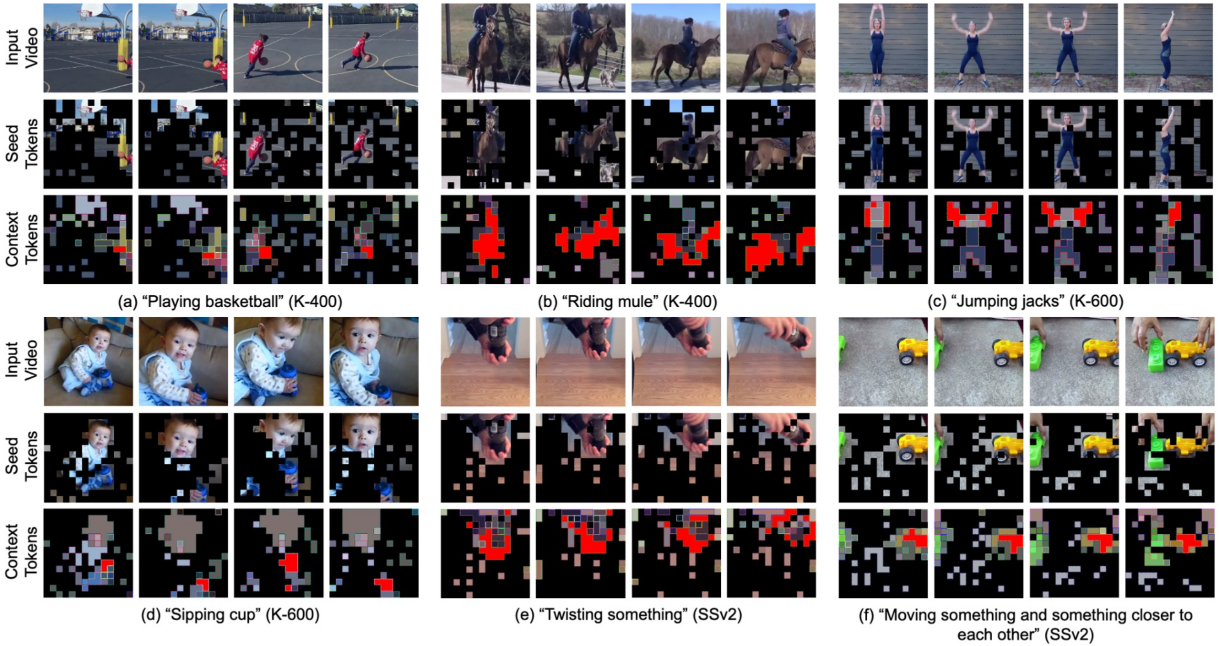
k	HMDB	UCF	SSv2	All
16	63.1	89.3	9.1	53.8
32	63.6	89.9	9.4	54.3
64	**63.7**	90.1	9.7	**54.5**
96	63.4	**90.2**	**9.9**	**54.5**
128	62.8	90.1	**9.9**	54.3

- **Text prompting conditioned on context tokens** is the most effective prompting design

Case	Use context tokens?	HMDB-51	UCF-101	K-600	All (Δ)
Baseline		52.3 ± 0.2	78.9 ± 1.1	70.7 ± 0.8	67.3
(a) Learnable prompt vectors		52.4 ± 0.4	78.4 ± 1.3	70.6 ± 0.7	67.1 (-0.2)
(b) Video-conditional prompting		53.2 ± 0.8	80.4 ± 0.7	71.6 ± 0.9	68.4 ($+1.1$)
(c) Video-conditional prompting	✓	53.7 ± 0.7	80.4 ± 0.9	72.7 ± 0.5	68.9 ($+1.6$)
(d) Vision-text late-fusion	✓	53.7 ± 0.7	79.0 ± 0.7	70.9 ± 0.6	67.9 ($+0.6$)

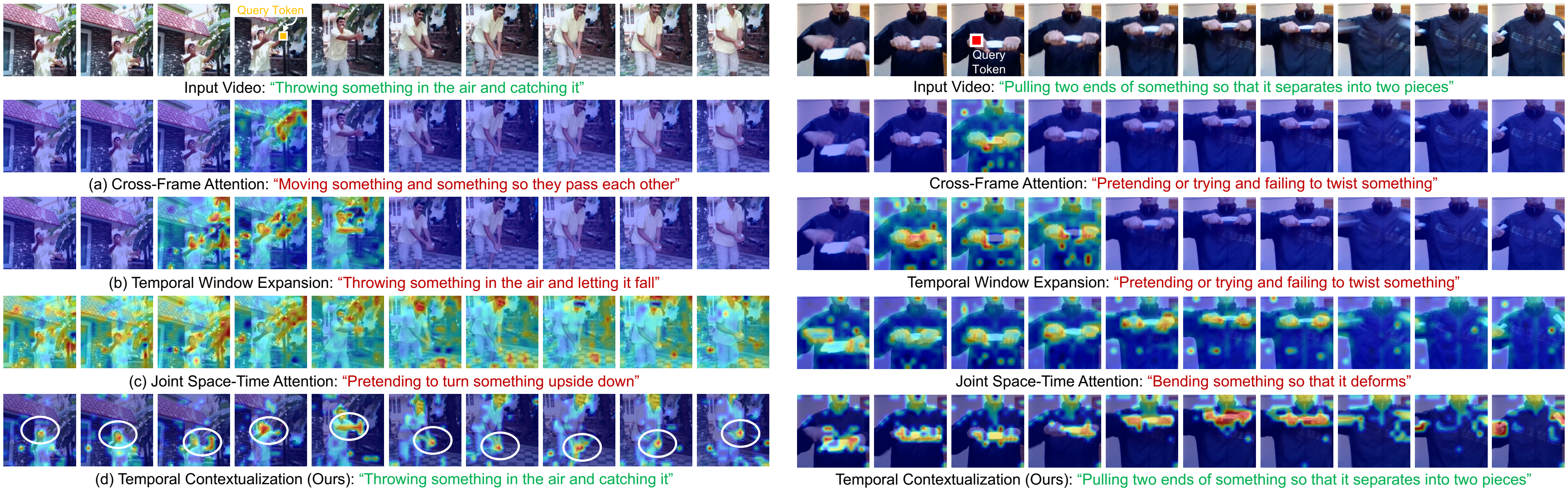
Visualizations

- Seed & context token visualization
  - **Seed tokens** mainly consist of patch tokens from the most informative regions in each frame
  - **Context tokens** successfully track and summarize a specific object or part throughout the video



Visualizations

- Attention visualization
  - **TC-CLIP correctly** predicts with **temporal consistency**
  - All other approaches fail to capture long-term temporal dynamics



Thank you

https://github.com/naver-ai/tc-clip

minji@snu.ac.kr
taekyung.k@navercorp.com

ECCV

ComputerVisionLab
Seoul National University

NAVER AI LAB

EUROPEAN
CONFERENCE
ON COMPUTER
VISION