Leveraging Temporal Contextualization for Video Action Recognition

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

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

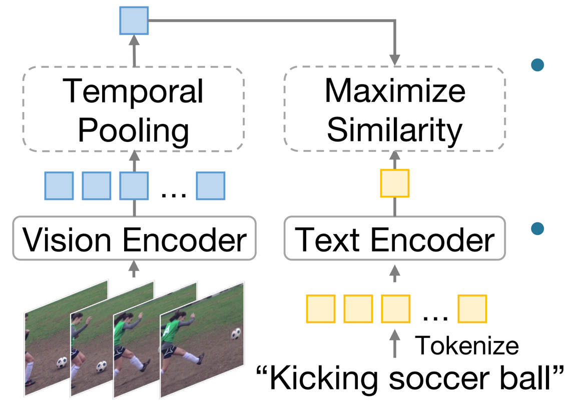
1 ComputerVisionLab Seoul National University 2 NAVER AI LAB

Paper Code

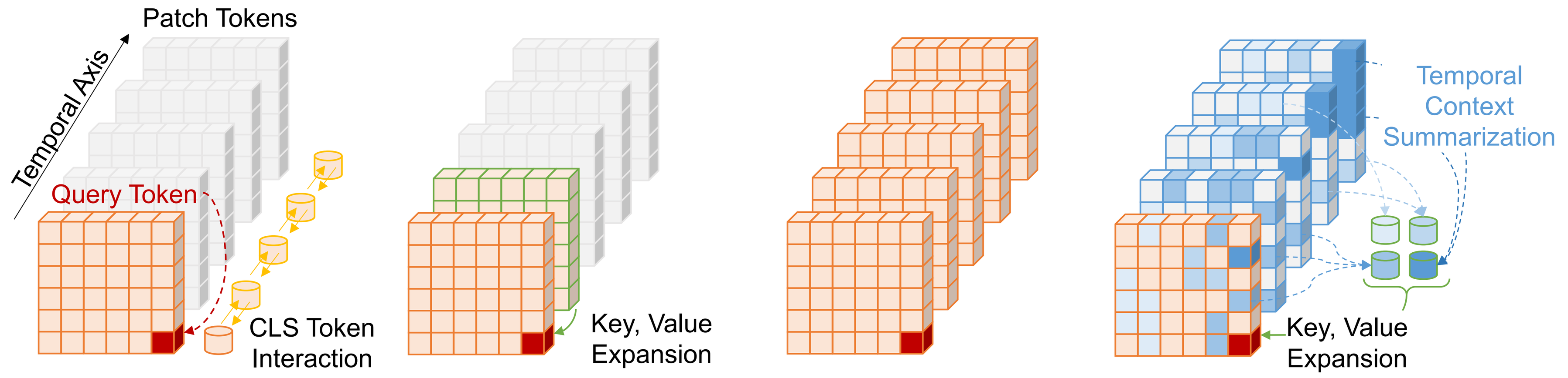
(See our paper for more results!)

Background

- Tuning **CLIP for video recognition** enables open-vocab generalization without expensive video-text pretraining.
- A naïve baseline: **frame-wise attention**
  → Limitation: **no token interactions** in the **temporal axis**
- To consider temporal cues, prior works additionally incorporate **reference tokens**: $\mathbf{z}_t^l = f_{\theta_v}^l(\mathbf{z}_t^{l-1}, \mathbf{s}^{l-1})$
  t-th frame patch tokens reference tokens

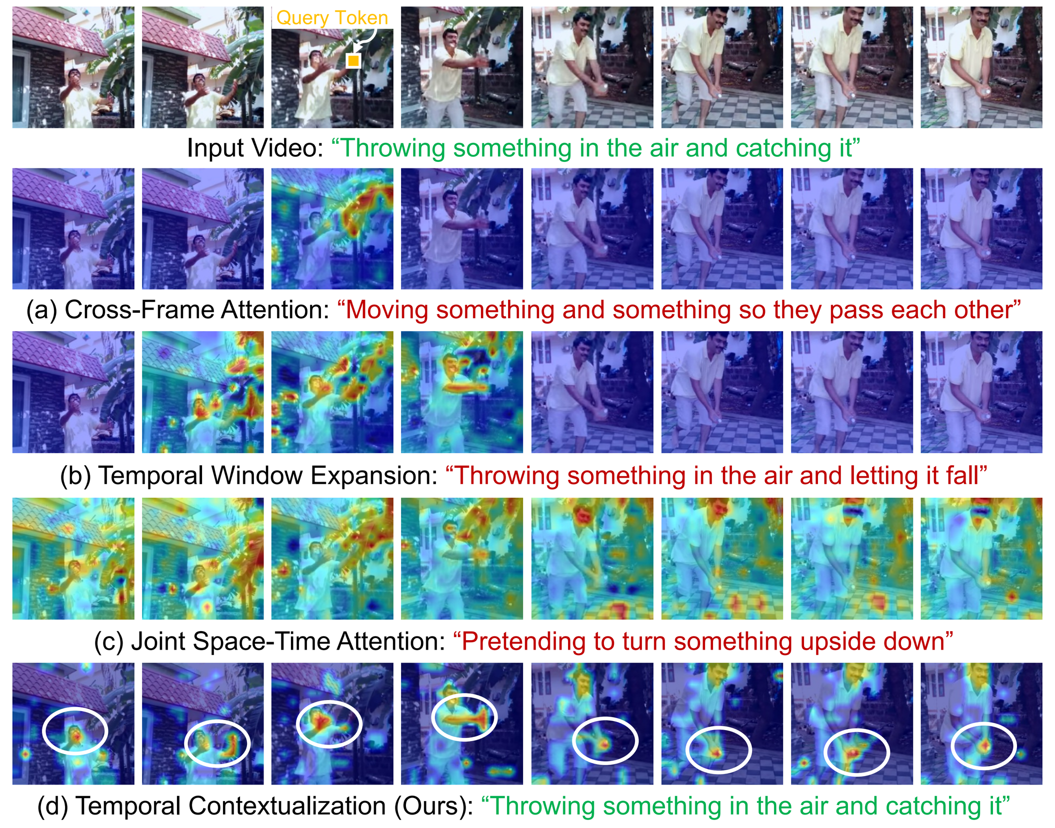


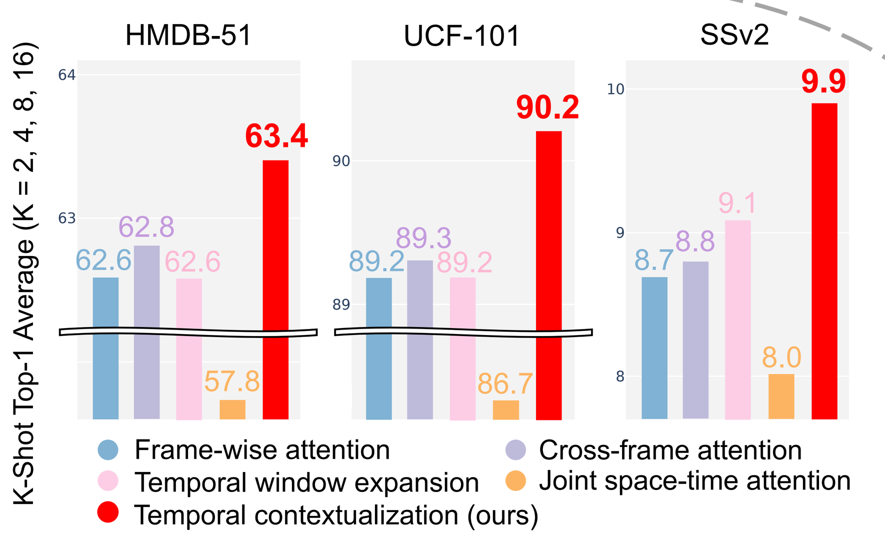
Problem: *Insufficient* Token Interactions in Temporal Modeling



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

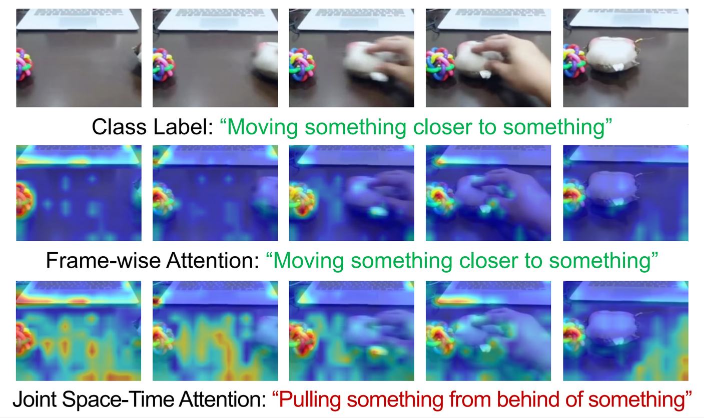
Reference Tokens	CLS tokens from all frames	Patch tokens from adjacent frames	Patch tokens from all frames	Context tokens
	Short-range token interactions		**Extrapolation challenge**	Deliver **global** information
	Fail to capture essential video information		Costly / Suboptimal ※	Maintain CLIP's **effective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





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

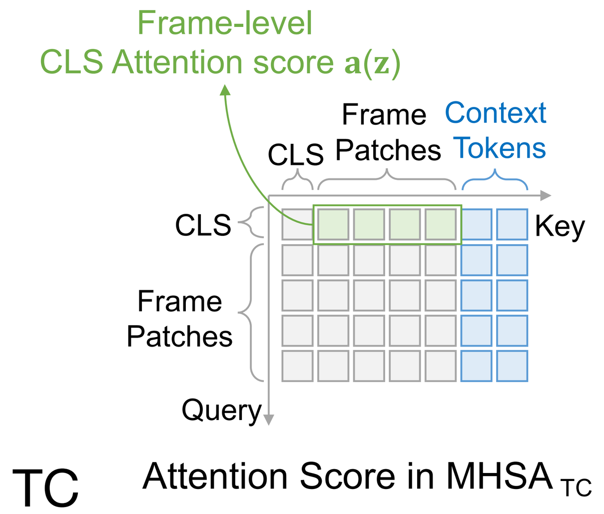
※ Pitfall of Joint Space-Time Attention

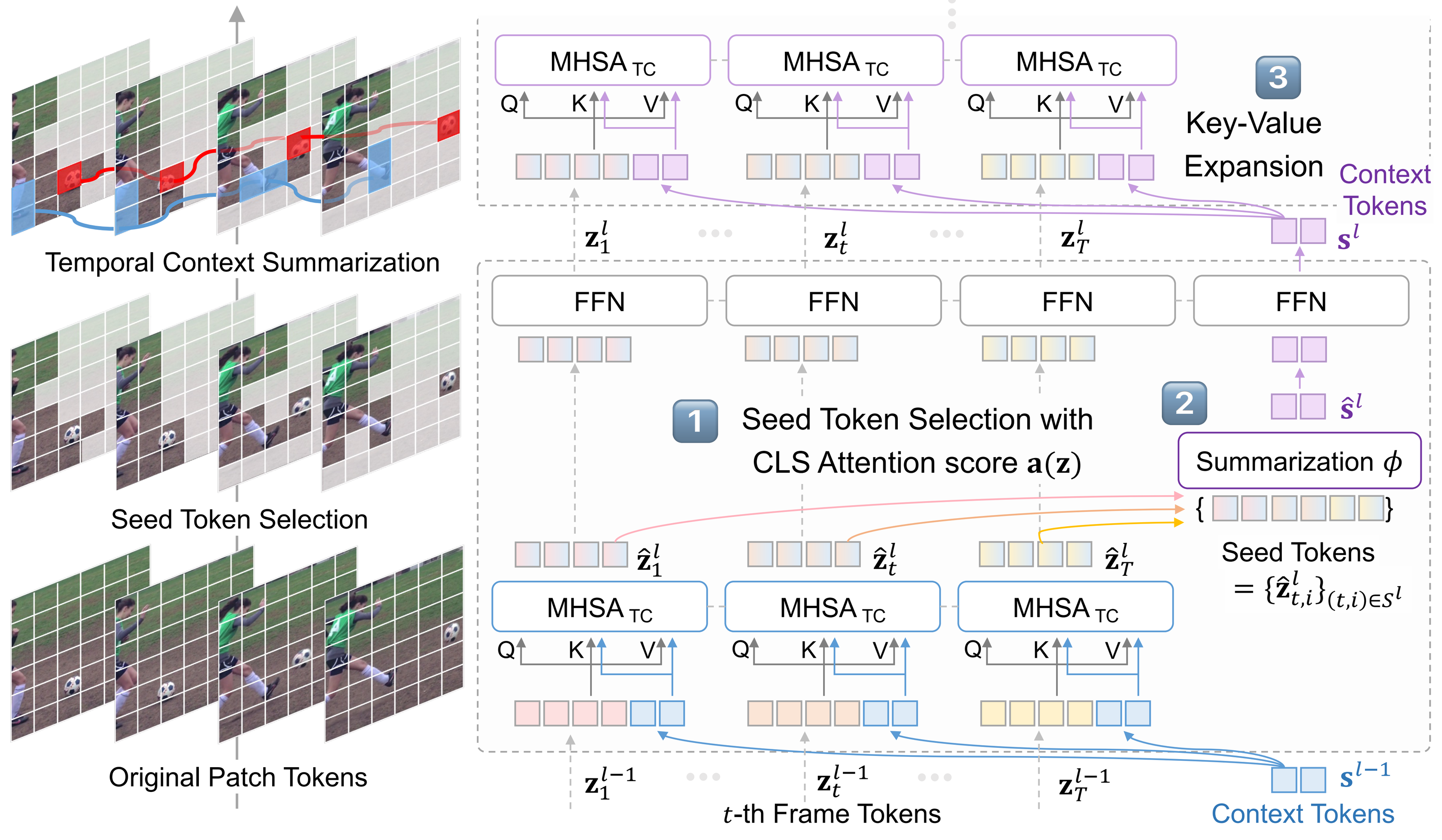


A naïve extension of CLIP's temporal sequence length **degrades attention quality** because it wasn't trained on long sequences.

Temporal Contextualization (TC)

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





- **Three steps of TC**
  1. **Seed token** selection in each frame
  2. Spatio-temporal context **summarization**
  3. **Temporal context infusion** to all tokens by **expanding key-value pairs**:

KV of context tokens

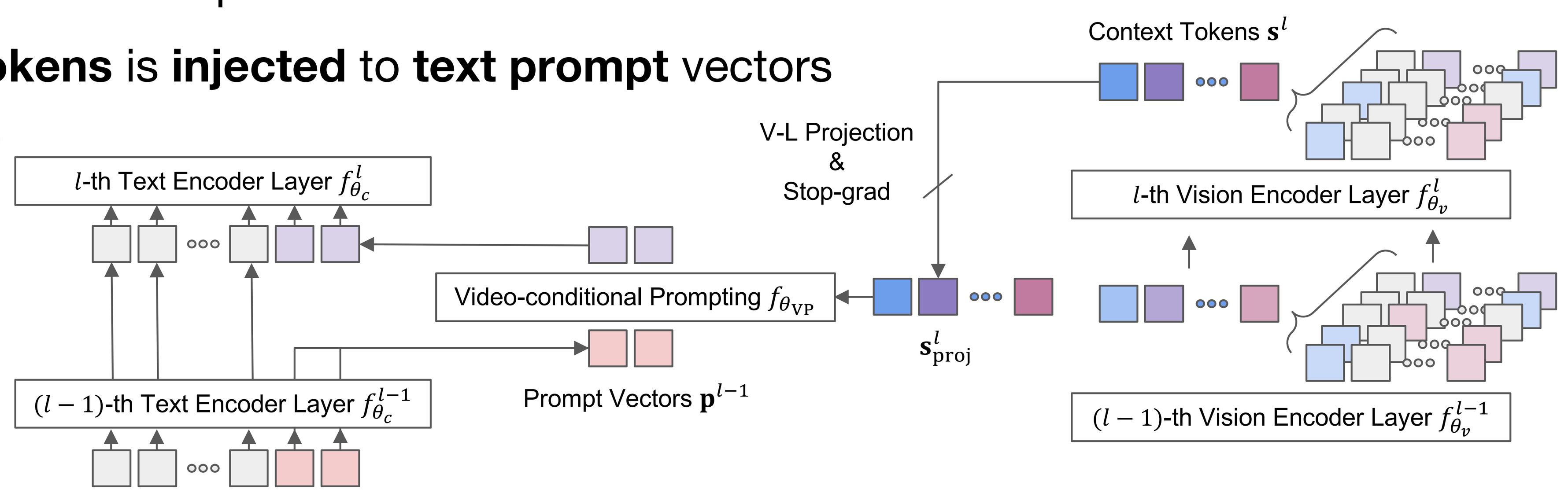
$$\mathrm{Attention}_{\mathrm{TC}}(\mathbf{z}_t, \mathbf{s}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{\mathbf{z}_t}[\mathbf{K}_{\mathbf{z}_t}|\mathbf{K}_{\mathbf{s}}]^{\top}}{\sqrt{d}} + \mathbf{B}\right)[\mathbf{V}_{\mathbf{z}_t}|\mathbf{V}_{\mathbf{s}}]$$

QKV of *t*-th frame patch tokens

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

Video-conditional Prompting (VP)

- VP generates **instance-level prompts** to compensate for the **lack of textual semantics**.
- **Video information from context tokens** is **injected** to **text prompt** vectors using a cross-attention mechanism.



Experiments

SOTA performance in **zero/few-shot, base2novel, fully-supervised** video recognition

- Results on **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**

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

	Without weight-space ensembling				With weight-space ensembling			
Case	HMDB-51	UCF-101	K-600	All (Δ)	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)

- **Context-token-conditional text prompting** is effective.

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)

Temporal Contextualization enhances **CLIP's video understanding** capability by **infusing global information** within the encoding process.