ICLR
International Conference On Learning Representations

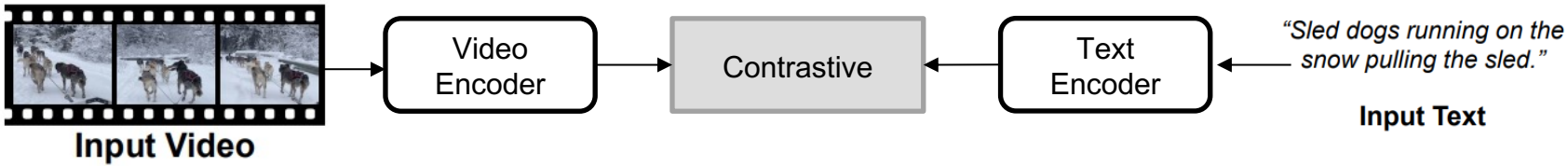
Structured Video-Language Modeling with Temporal Grouping and Spatial Grounding

Yuanhao Xiong[1,3], Long Zhao[1], Boqing Gong[1], Ming-Hsuan Yang[1], Florian Schroff[1], Ting Liu[1], Cho-Jui Hsieh[2,3], Liangzhe Yuan[1]

[1] Google Research [2] Google [3] UCLA

Motivation

- Most existing works merely focus on learning holistic global features.



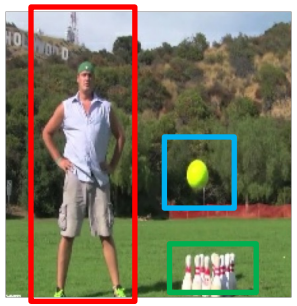
- However, fine-grained structures such as the correspondences between regions in a video and nouns in a caption and distinction between temporal frames have been proved important for downstream tasks.

Motivation

- Fine-grained structures

1.0	0.96	0.77	0.9
0.96	1.0	0.88	0.94
0.77	0.88	1.0	0.86
0.9	0.94	0.86	1.0

Similar scores between different frames.

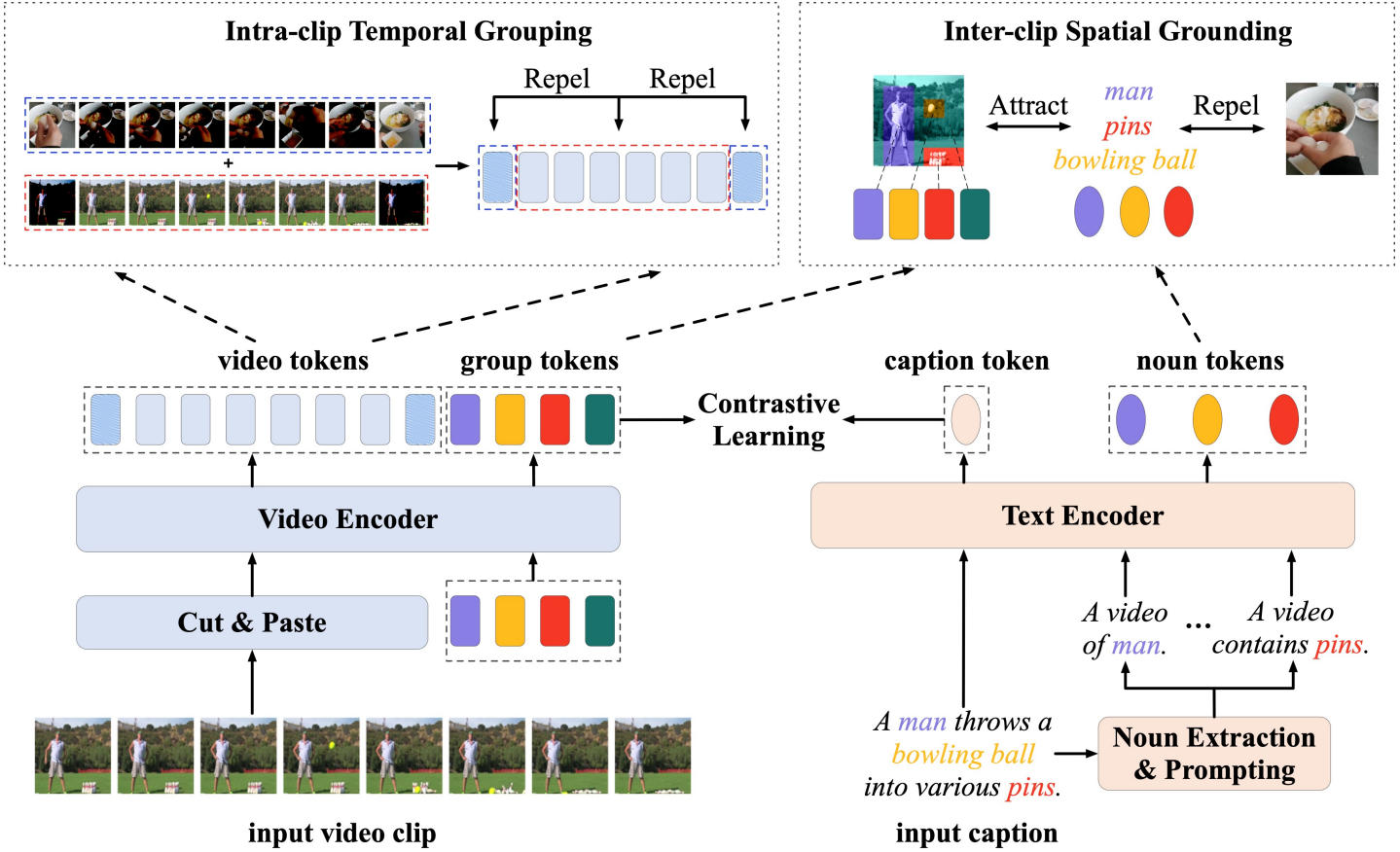


A ***man*** throws a ***bowling ball*** into various ***pins***.

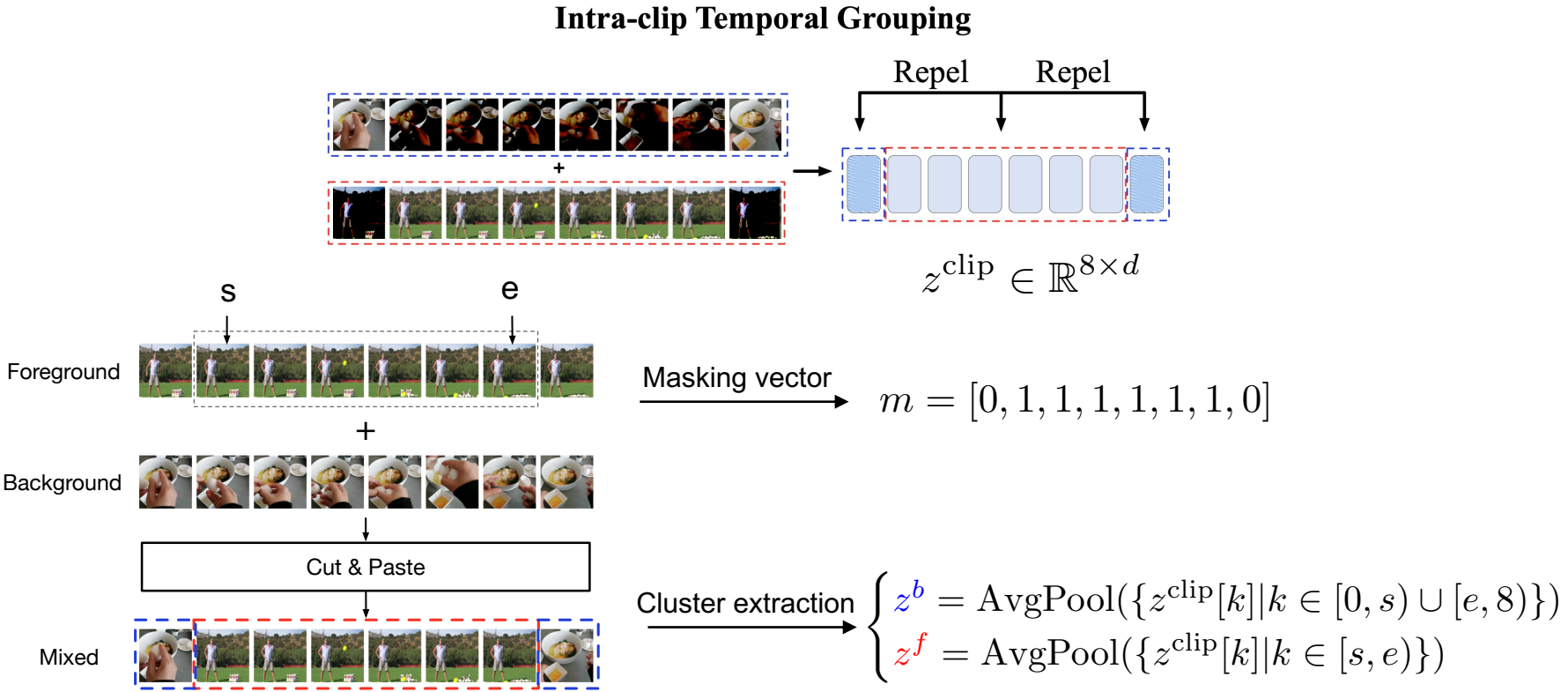
Correspondence between regions and nouns

- Is it possible to integrate these fine-grained structures into training?

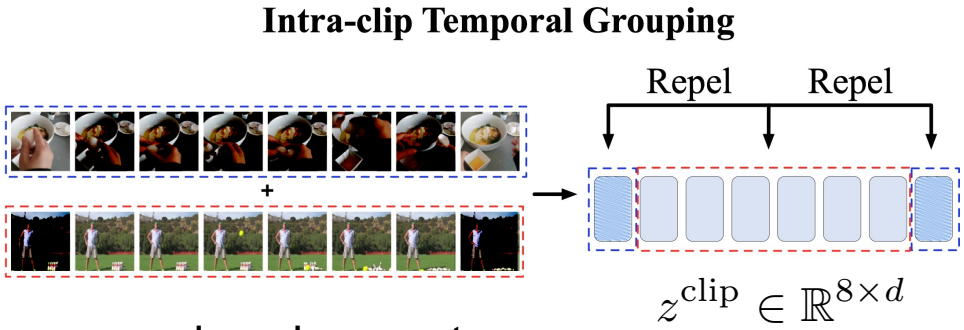
S-ViLM (**S**tructured **Vi**deo-**L**anguage **M**odeling)



Temporal Grouping



Temporal Grouping



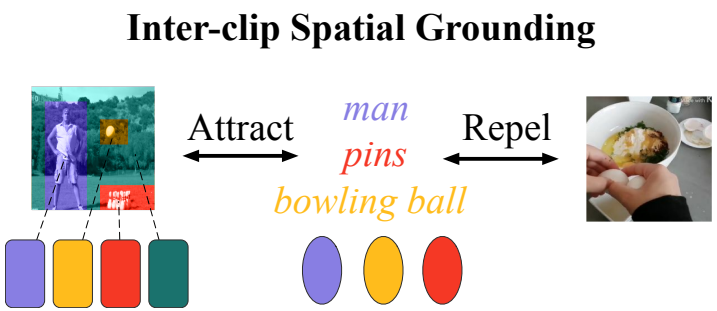
- Background-foreground assignment

$$a = \text{softmax}(\langle z^{\text{clip}}, [z^b; z^f]^T \rangle / \tau) \in \mathbb{R}^{8\times 2}$$

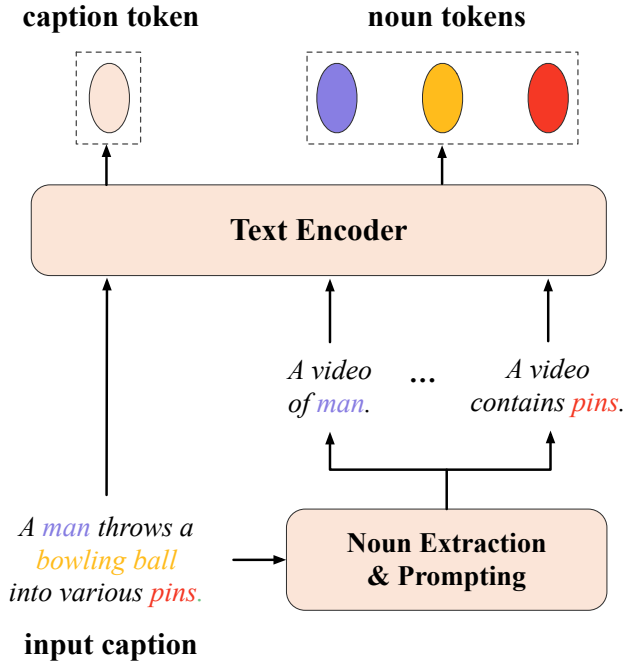
- Grouping as a binary classification problem

$$\mathcal{L}_{\text{t}} = \frac{1}{B}\sum_{i}^{B} \ell_{\text{BCE}}(a_i, \text{One-hot}(m_i))$$

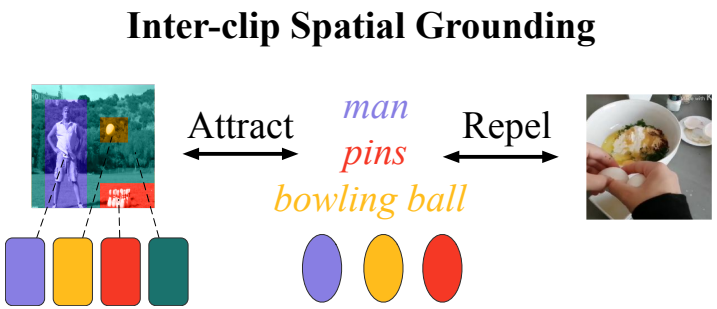
Spatial Grounding

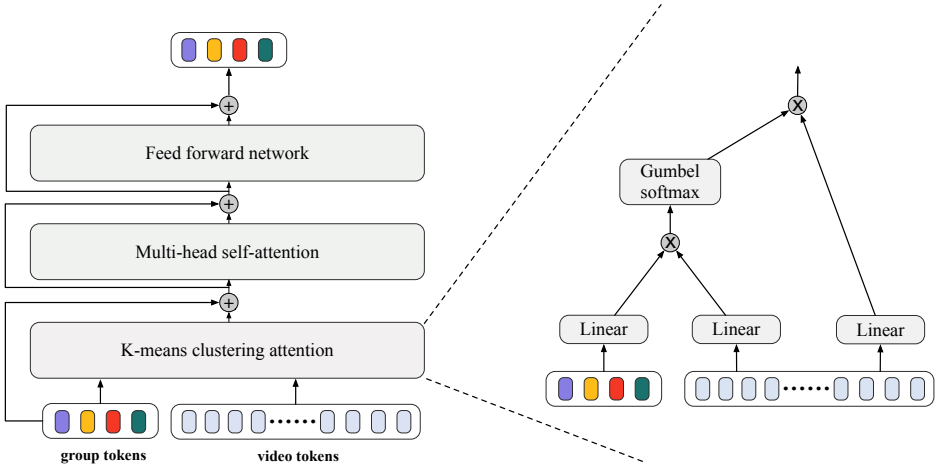


- Noun token extraction
  - Extract nouns from the caption
  - Prompt with templates
  - Feed into the text encoder to obtain noun tokens



Spatial Grounding

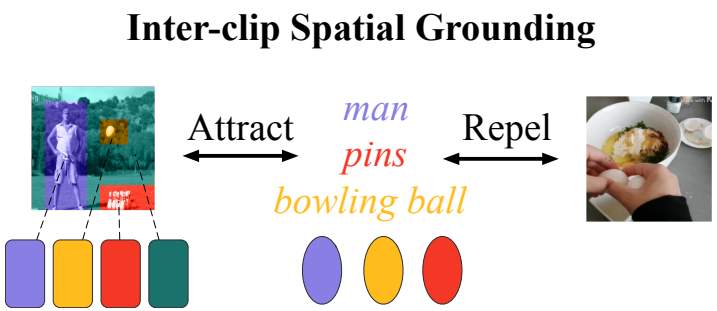


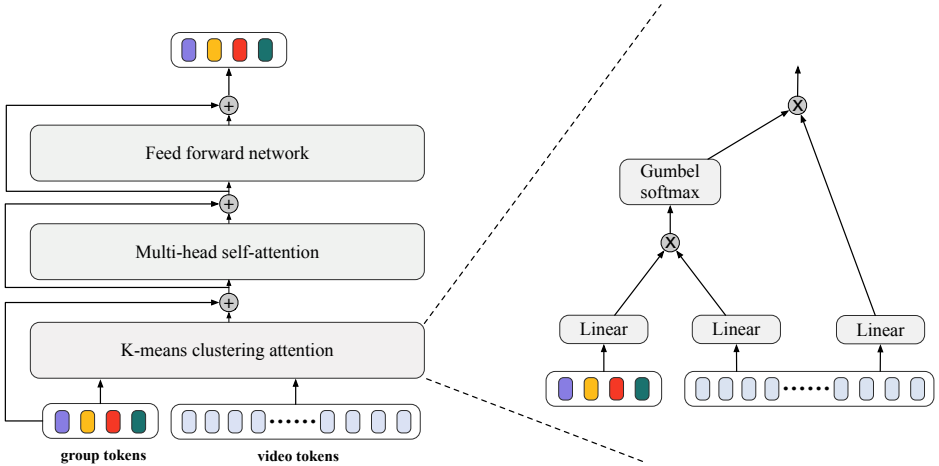


Grouping block

- Region token extraction
  - Region tokens are extracted by group tokens
  - They aggregate semantically similar video patches into clusters.

Spatial Grounding





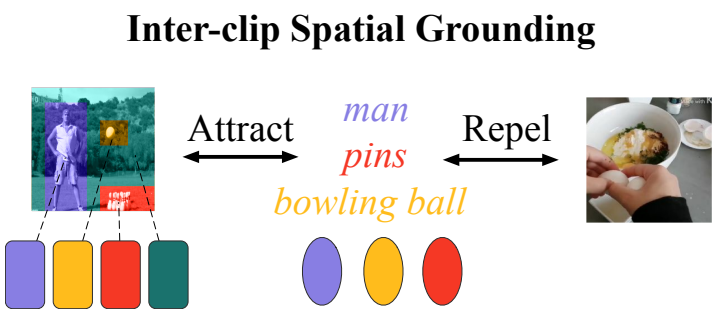
Grouping block

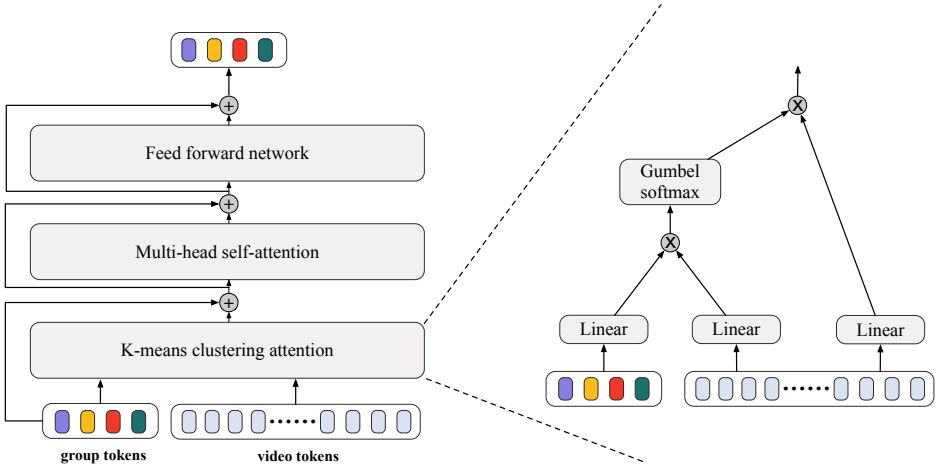
- Grounding similarity
  - Intuitively, it encourages each noun to be grounded to one or a few group tokens.

$$G(v,c) = \frac{1}{K}\sum_{k=1}^{K}\left\langle f^{n_k}, \sum_{m=1}^{N_g}\frac{\exp(\langle f^{n_k}, f^{g_m}\rangle)}{\sum_{i=1}^{N_g}\exp(\langle f^{n_k}, f^{g_i}\rangle)}\cdot f^{g_m}\right\rangle$$

weighted group token

Spatial Grounding



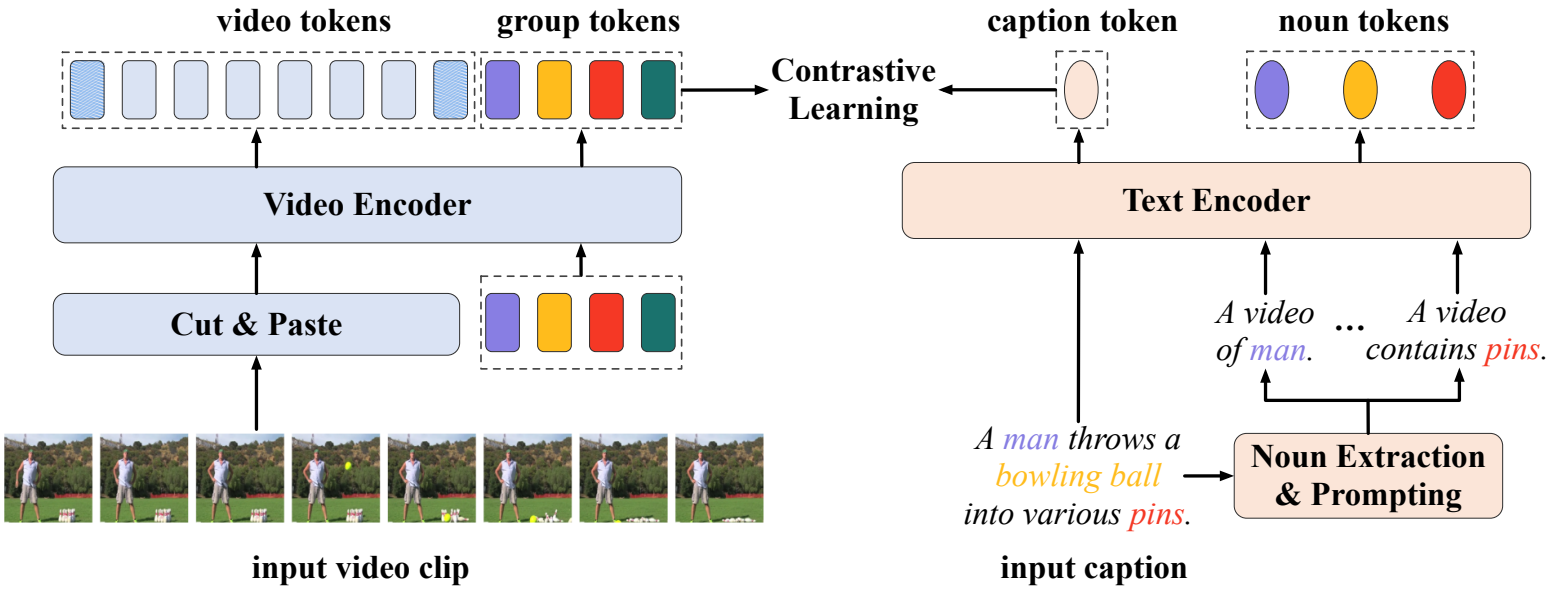


Grouping block

- Grounding loss

$$\begin{cases} \mathcal{L}_{\mathrm{g}}^{v\to c} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(G(v_i,c_i)/\tau)}{\sum_{j=1}^{B}\exp(G(v_i,c_j)/\tau)} \\ \mathcal{L}_{\mathrm{g}}^{c\to v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(G(v_i,c_i)/\tau)}{\sum_{j=1}^{B}\exp(G(v_j,c_i)/\tau)} \end{cases}$$

Global Contrastive Loss



Experimental Results: Text-Video Retrieval

Method	Video Encoder Input	R@1	R@5	R@10	MedR
MIL-NCE (Miech et al., 2019)	Raw Videos	9.9	24.0	32.4	29.6
VATT (Akbari et al., 2021)	Raw Videos	-	-	29.7	49.0
VideoCLIP (Xu et al., 2021b)	S3D	10.4	22.2	30.0	-
SupportSet (Patrick et al., 2020)	R(2+1)D-34	12.7	27.5	36.2	24.0
Frozen (Bain et al., 2021)	Raw Videos	18.7	39.5	51.6	10.0
AVLnet (Rouditchenko et al., 2021)	ResNeXt-101	19.6	40.8	50.7	9.0
DemoVLP (Cai et al., 2022)	Raw Videos	24.0	44.0	52.6	8.0
ALPRO (Li et al., 2022)	Raw Videos	24.1	44.7	55.4	8.0
MCQ (Ge et al., 2022a)	Raw Videos	26.0	46.4	56.4	7.0
VCC (Nagrani et al., 2022)	Raw Videos	18.9	37.5	47.1	-
S-ViLM	Raw Videos	**28.6**	**53.6**	**65.1**	**5.0**
UniVL (Luo et al., 2020)	S3D	21.2	49.6	63.1	6.0
MMT (Gabeur et al., 2020)	S3D	26.6	57.1	69.6	4.0
ClipBERT (Lei et al., 2021)	Raw Videos	22.0	46.8	59.9	6.0
AVLnet (Rouditchenko et al., 2021)	ResNeXt-101	27.1	55.6	66.6	4.0
SupportSet (Patrick et al., 2020)	R(2+1)D-34	30.1	58.5	69.3	3.0
VideoCLIP (Xu et al., 2021b)	S3D	30.9	55.4	66.8	-
Frozen (Bain et al., 2021)	Raw Videos	31.0	59.5	70.5	3.0
DemoVLP (Cai et al., 2022)	Raw Videos	36.0	61.0	71.8	3.0
ALPRO (Li et al., 2022)	Raw Videos	33.9	60.7	73.2	3.0
MCQ (Ge et al., 2022a)	Raw Videos	37.6	64.8	75.1	3.0
VIOLETv2 (Fu et al., 2023)	Raw Videos	37.2	64.8	75.8	-
All-in-One (Wang et al., 2023b)	Raw Videos	37.1	**66.7**	75.9	-
VCC (Nagrani et al., 2022)	Raw Videos	35.0	63.1	75.1	-
S-ViLM	Raw Videos	**38.4**	65.7	**76.3**	**2.0**

Text-video retrieval evaluation on MSR-VTT.

Experimental Results: VQA & Action Recognition & TAL

Method	MSRVTT-QA	MSVD-QA
HGA (Jiang & Han, 2020)	35.5	34.7
QUEST (Jiang et al., 2020)	34.6	36.1
HCRN (Le et al., 2020)	35.6	36.1
ClipBERT (Lei et al., 2021)	37.4	-
SSML (Amrani et al., 2021)	35.1	35.1
CoMVT (Seo et al., 2021)	39.5	42.6
DemoVLP (Cai et al., 2022)	38.3	39.5
ALPRO (Li et al., 2022)	42.1	45.9
S-ViLM	**43.5**	**46.4**

VQA

Method	Modal	UCF101		HMDB51	
		Lin	FT	Lin	FT
CoCLR (Han et al., 2020)	OF	77.8	90.6	52.4	62.9
MVCGC (Huo et al., 2021)	MV	78.0	90.8	53.0	63.4
XDC_R (Alwassel et al., 2020)	A	80.7	88.8	49.9	61.2
XDC_K (Alwassel et al., 2020)	A	85.3	91.5	56.0	63.1
MIL-NCE (Miech et al., 2019)	T	83.4	89.1	54.8	59.2
Frozen (Bain et al., 2021)	T	87.8	89.8	61.3	66.3
VATT (Akbari et al., 2021)	A, T	89.2	-	63.3	-
ELO (Piergiovanni et al., 2020)	A, OF	-	93.8	64.5	67.4
MMV (Alayrac et al., 2020)	A	77.1	-	53.6	-
MMV (Alayrac et al., 2020)	T	86.8	-	55.1	-
MMV (Alayrac et al., 2020)	A, T	91.8	95.2	67.1	75.0
MCQ (Ge et al., 2022a)	T	89.1	92.3	65.8	69.8
S-ViLM	T	**94.8**	**96.5**	**70.0**	**76.9**

Action Recognition

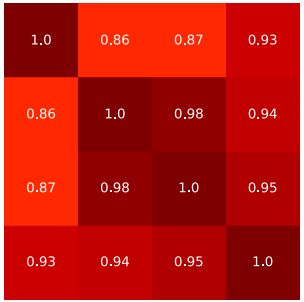
Method	TAL Task (G-TAD)			
	mAP@0.5	@0.75	@0.95	Avg
CoCLR (Han et al., 2020)	47.9	32.3	7.3	31.9
XDC (Alwassel et al., 2020)	48.4	32.6	7.6	32.3
MoCo-v2 (Chen et al., 2020a)	46.6	30.7	6.3	30.3
VideoMoCo (Pan et al., 2021)	47.8	32.1	7.0	31.7
RSPNet (Chen et al., 2021)	47.1	31.2	7.1	30.9
AoT (Wei et al., 2018)	44.1	28.9	5.9	28.8
SpeedNet (Benaim et al., 2020)	44.5	29.5	6.1	29.4
PAL (Zhang et al., 2022)	50.7	35.5	8.7	34.6
TAC (Xu et al., 2020)	48.5	32.9	7.2	32.5
BSP (Xu et al., 2021c)	50.9	35.6	8.0	34.8
LoFi (Xu et al., 2021d)	50.4	35.4	8.9	34.4
TSP (Alwassel et al., 2021)	51.3	**37.1**	9.3	**35.8**
S-ViLM	**51.7**	36.4	**9.7**	35.6

Temporal Action Localization

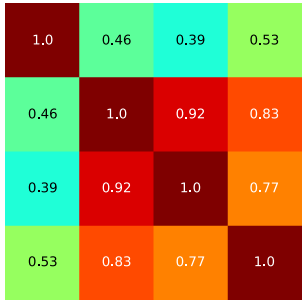
Visualization

Intra-clip Temporal Grouping

Caption: A man was driving a tractor for land reclamation.

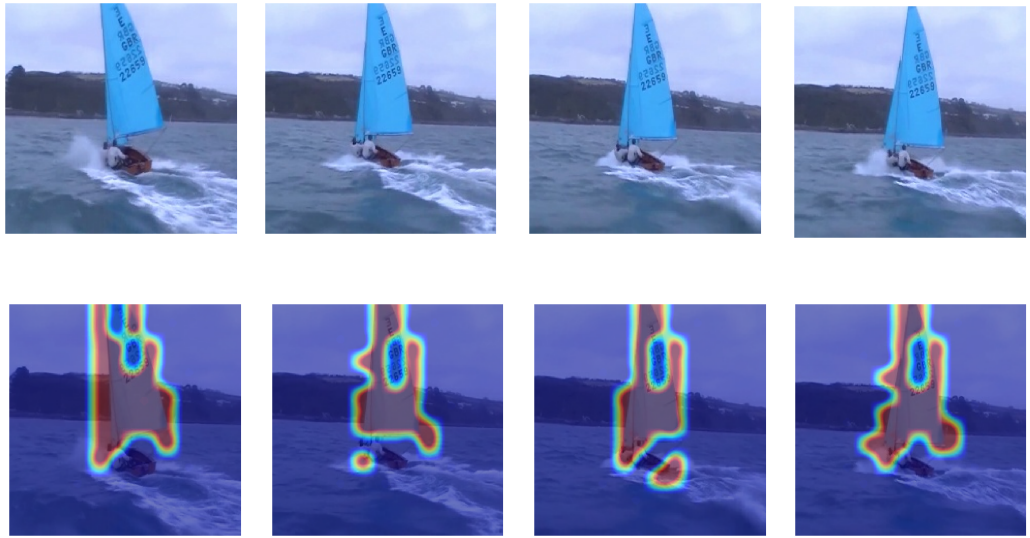


baseline



temporal-aware

Inter-clip Spatial Grounding



Caption: The boat makes a turn to the right.

Summary

- We propose S-ViLM, a dual-encoder video-language modeling framework, making use of structured video-caption interactions
  - Temporal grouping to learn temporal-aware features
  - Spatial grounding to align regions and noun phrases

- Experimental results have demonstrated the effectiveness of S-ViLM on four downstream tasks, including ***text-video retrieval***, ***video question answering***, ***video action recognition***, and ***temporal action localization***