ICML
International Conference
On Machine Learning

CLIP Tricks You:

Training-free Token Pruning for Efficient Pixel Grounding in Large Vision-Language Models

Sangin Lee[1], Yukyung Choi[1,2]

[1] Dept. of AI and Robotics, Sejong Univ.
[2] AI and Robotics Institute (AIRI), Sejong Univ.

Background: Large Vision-Language Model (LVLM)

Background: Large Vision-Language Model (LVLM)

Background

Image Captioning

Describe this photo in detail.

→

Large Vision-Language Model

→

A young boy wearing a soccer jersey and blue shorts is kicking a soccer ball across a grassy field ···

Background

Visual Question Answering (VQA)

How many players are visible?

Large Vision-Language Model

There are **2** players in the image.

Background

Optical Character Recognition (OCR)

What text is written on the jersey?

Large Vision-Language Model

"CHERRY HILL"

Background

Referring Expression Segmentation

Can you segment the ball near the boy in blue shorts?

Large Vision-Language Model

Motivation

Image Captioning

Describe this photo in detail.

Visual Question Answering (VQA)

How many players are visible?

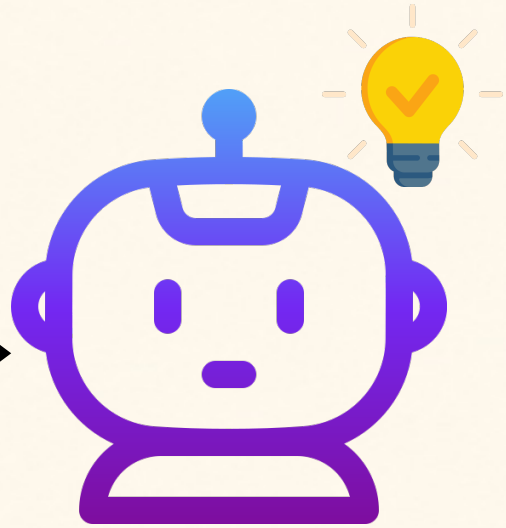
Optical Character Recognition (OCR)

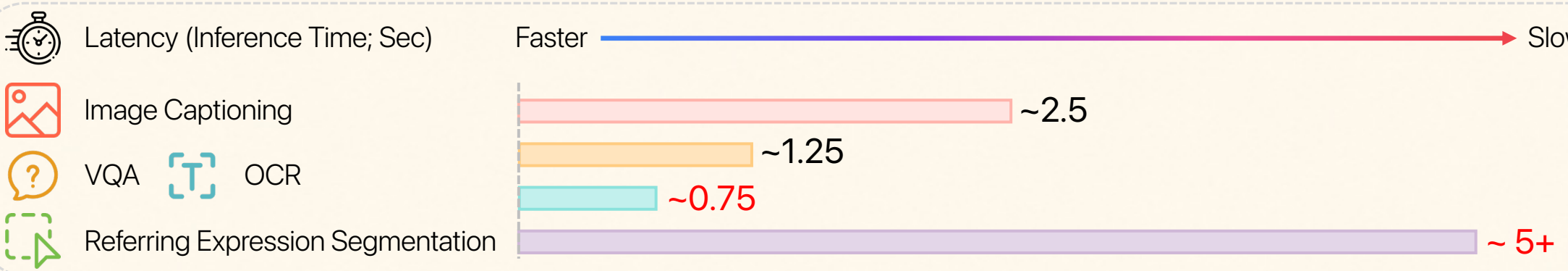
What text is written on the jersey?

Referring Expression Segmentation

Can you segment the ball near the boy in blue shorts?

Large Vision-Language Model





Motivation

Image Captioning

Describe this photo in detail.

Large Vision-Language Model

A young boy wearing a soccer jersey and blue shorts is kicking a soccer ball across a grassy field ···

 # Visual Token: 576
 # Text Token : 6~8

Motivation

Image Captioning

Describe this photo in detail.

→

Large Vision-Language Model

→

A young boy wearing a soccer jersey and blue shorts is kicking a soccer ball across a grassy field …

 # Visual Token: 2880
 # Text Token : 6~8

Motivation

Optical Character Recognition (OCR)

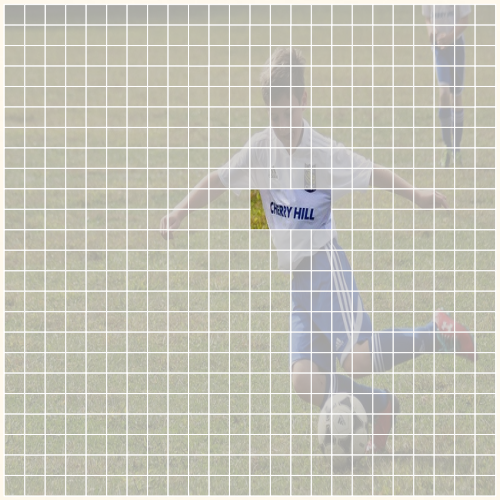
What text is written on the jersey?

Large Vision-Language Model

"CHERRY HILL"

 # Visual Token: 576+
 # Text Token : 8~10

Should All Visual Tokens Be Retained?



Optical Character Recognition (OCR)

What text is written on the jersey?

Large Vision-Language Model

"CHERRY HILL"

Should All Visual Tokens Be Retained?

Referring Expression Segmentation

Can you segment the ball
near the boy in blue shorts?

Please segment
the player wearing green cleats

Large Vision-Language Model

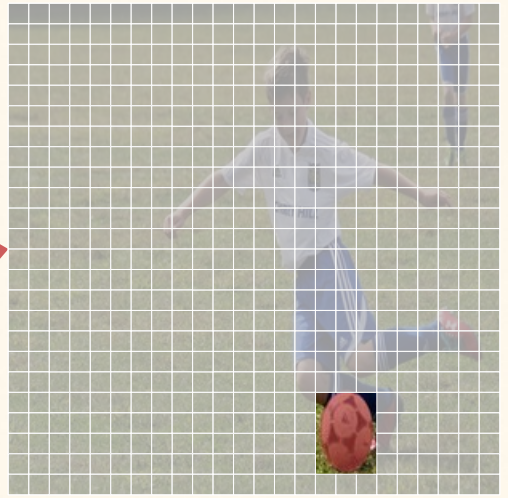
Should All Visual Tokens Be Retained?

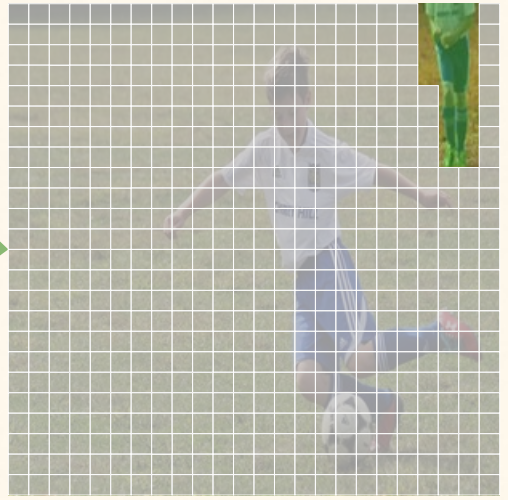
Referring Expression Segmentation

Can you segment the ball near the boy in blue shorts?

Please segment the player wearing green cleats

Large Vision-Language Model

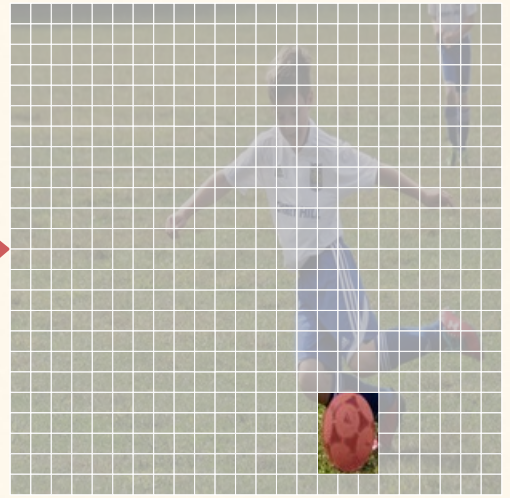


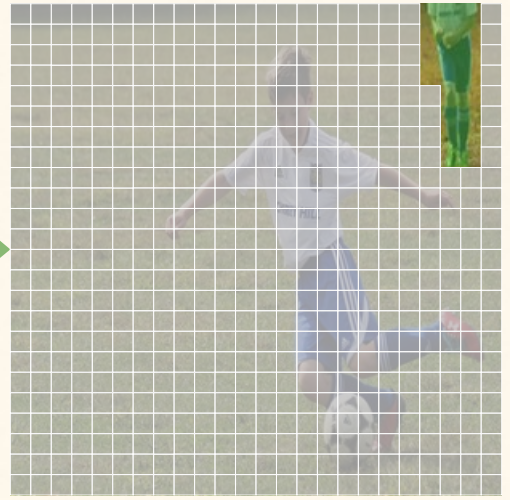


Should All Visual Tokens Be Retained?

Large Vision-Language Model

Token Pruning!





The Answer Lies in Token Pruning

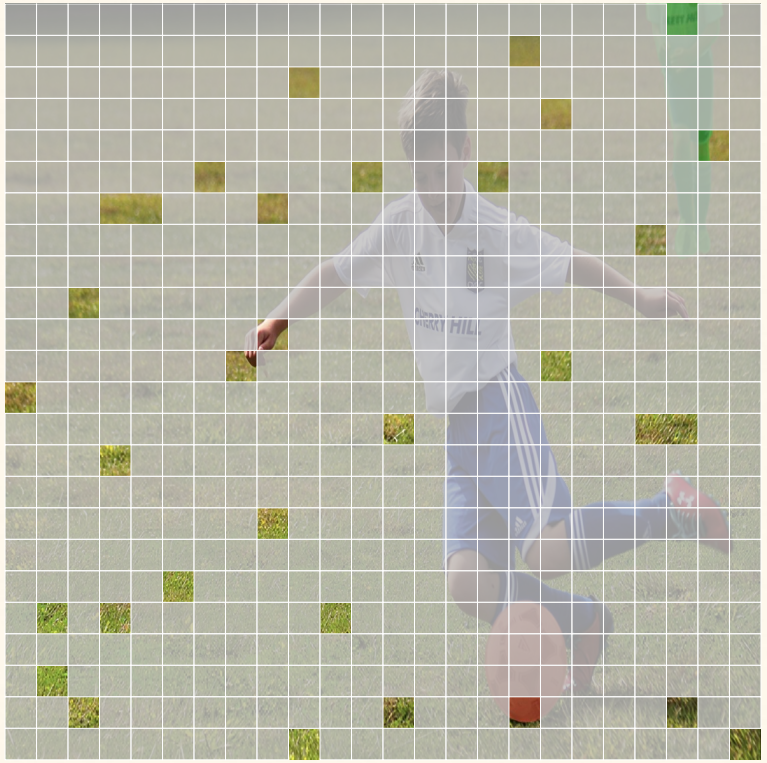
- Token pruning reduce redundant or less informative visual tokens.

The Answer Lies in Token Pruning

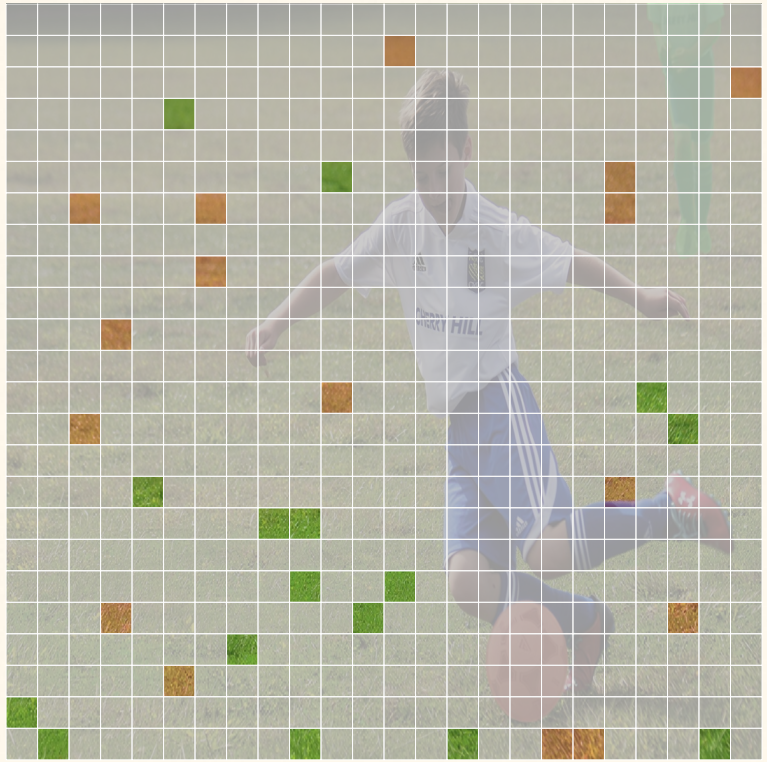
- Token pruning reduce redundant or less informative visual tokens.
- Token pruning has been actively studied and proven effective for image understanding tasks.

Is Token Pruning Always Effective?

- Token pruning reduce redundant or less informative visual tokens.
- Token pruning has been actively studied and proven effective for image understanding tasks.
- Yet, we find token pruning falls short in pixel grounding (e.g., referring expression segmentation).



LLaVA-PruMerge[1]

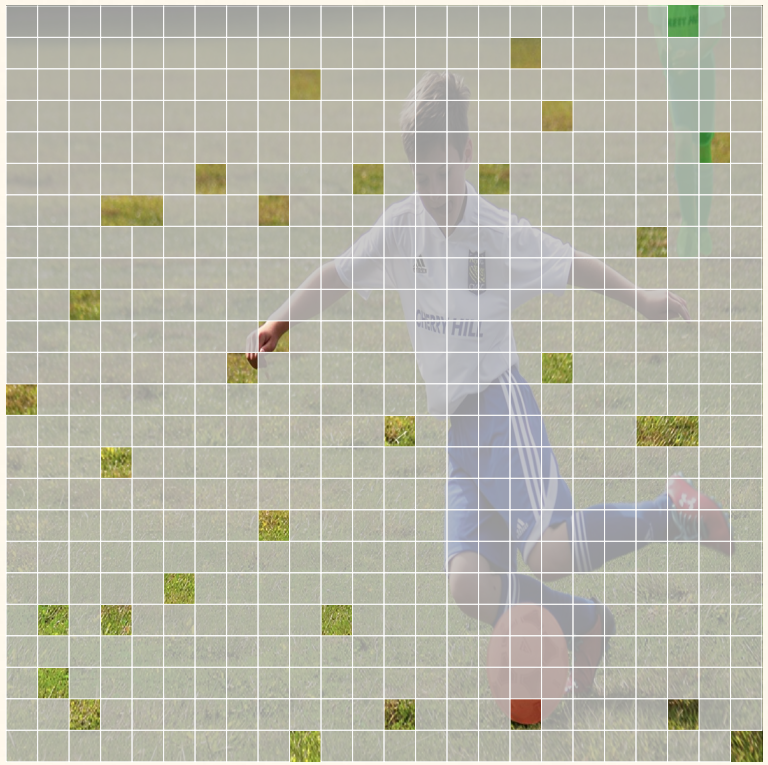


TRIM[2]

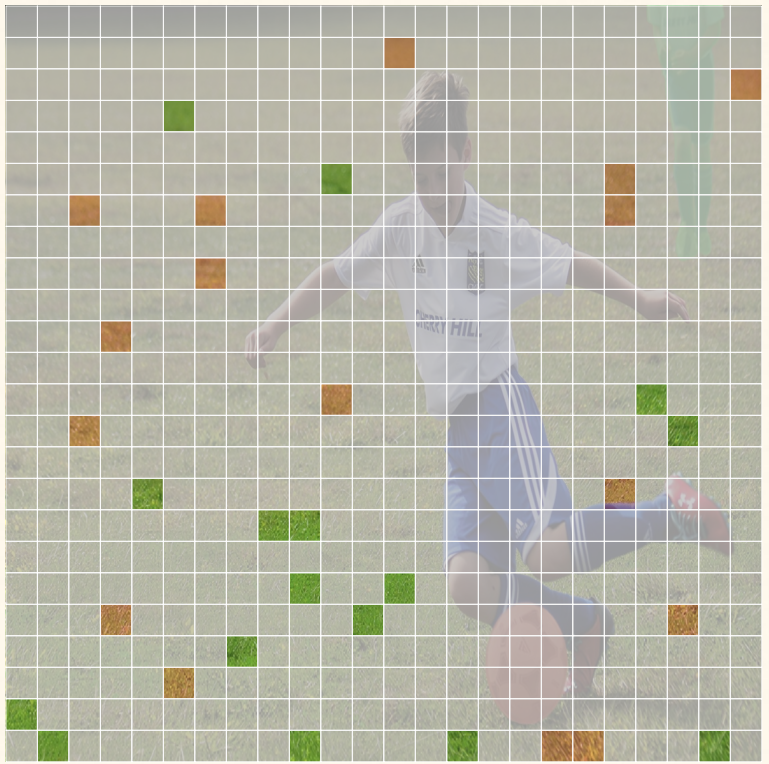
[1] Shang et al., LLaVA-PruMerge ···, ICCV, 2025.
[2] Song et al., Less is More ···, COLING, 2025.

Is Token Pruning Always Effective?

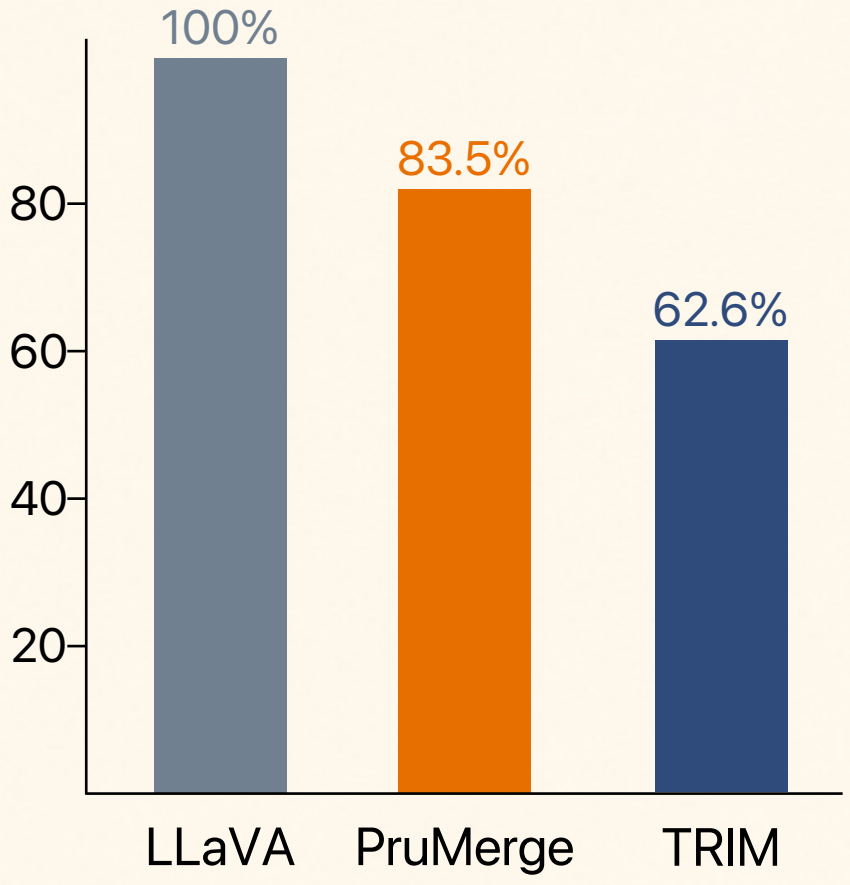
- Token pruning has been actively studied and proven effective for image understanding tasks.
- Yet, we find token pruning falls short in pixel grounding (e.g., referring expression segmentation).
- If visual tokens within the referent are pruned, performance degrades significantly.



LLaVA-PruMerge



TRIM



What Are We Missing? Text Awareness

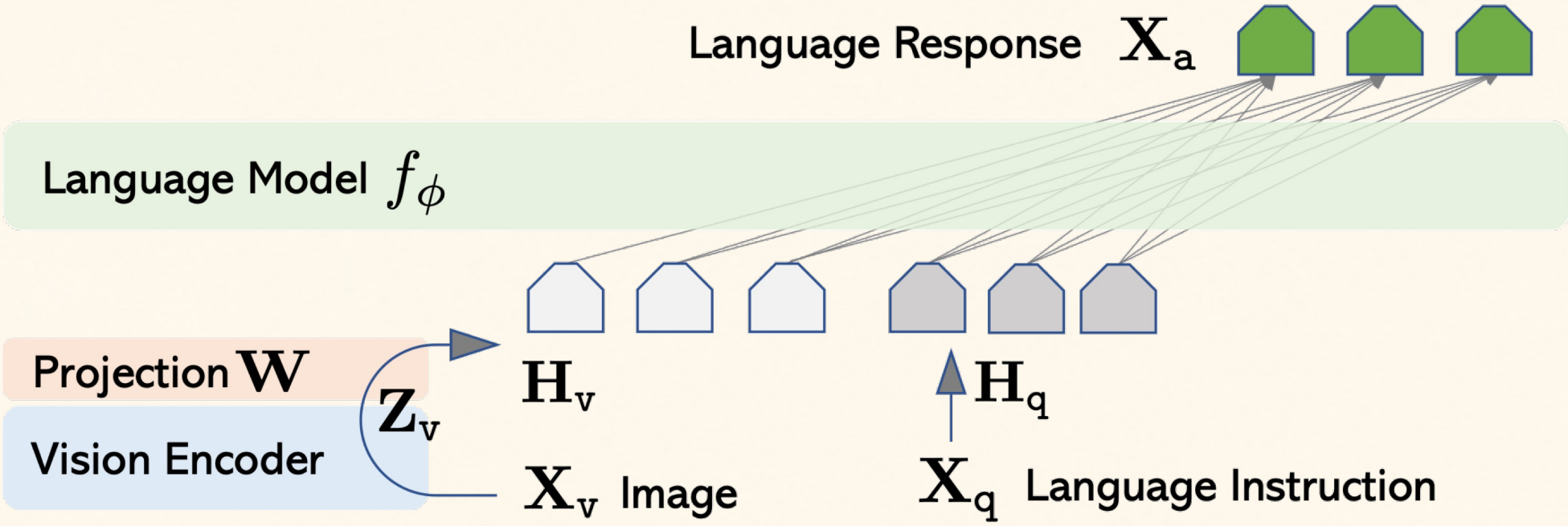
- Most token pruning methods are text-agnostic.

What Are We Missing? Text Awareness

- Most token pruning methods are text-agnostic.
- They primarily rely on visual information, such as global visual similarity.

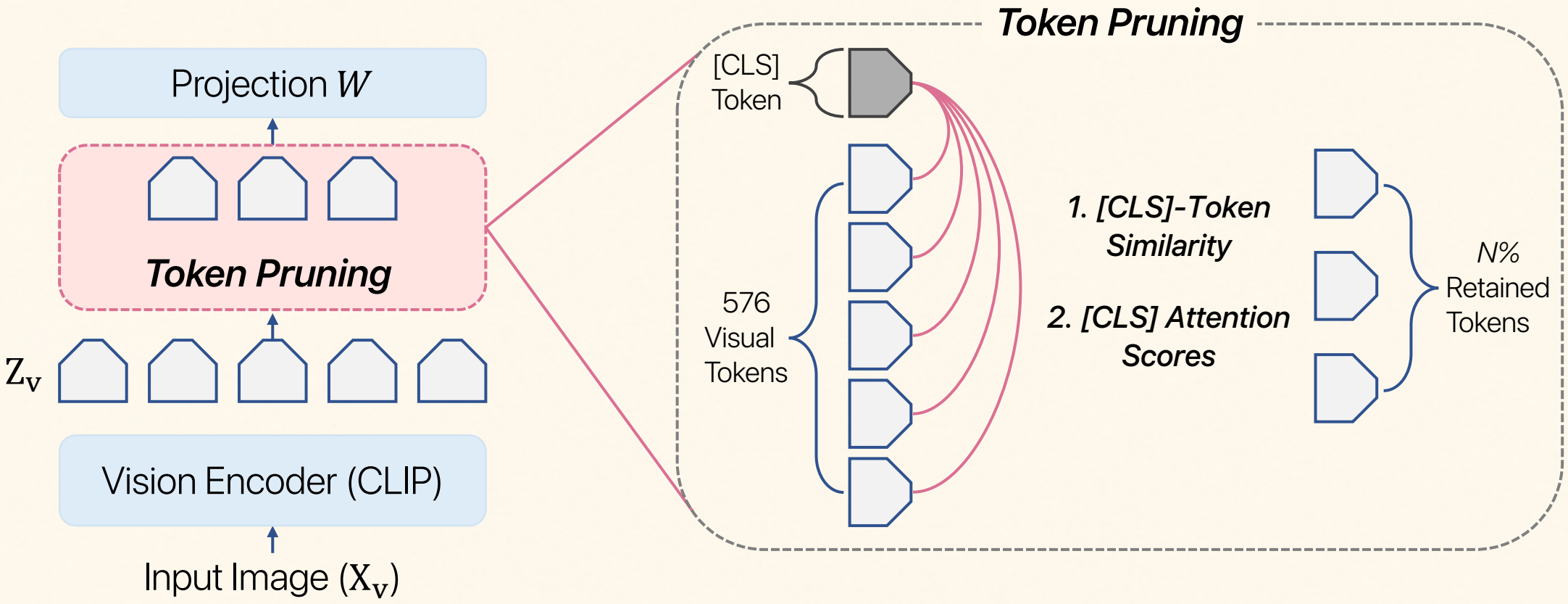
A Brief Look at LVLM – LLaVA[3]

- For an input image $\mathbf{X_v}$, vision encoder CLIP[4] generates 576 visual tokens ($\mathbf{Z_v}$).



[3] Liu et al., Visual Instruction Tuning, NeurIPS, 2023.
[4] Radford et al., Learning Transferable ···, ICML, 2021.

What Are We Missing? Text Awareness



What Are We Missing? Text Awareness

- Most token pruning methods are text-agnostic.
- They primarily rely on visual information, such as global visual similarity.
- As a result, they fail to adapt to text-specified targets.

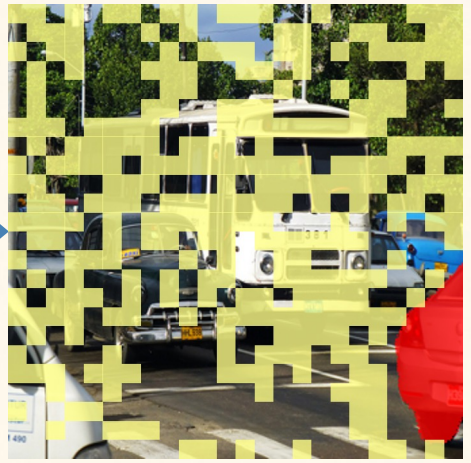
Input Text 1

Where is the blk car?
Please return a segmentation mask

Input Text 2

Can you highlight the
bottom right car in this image?

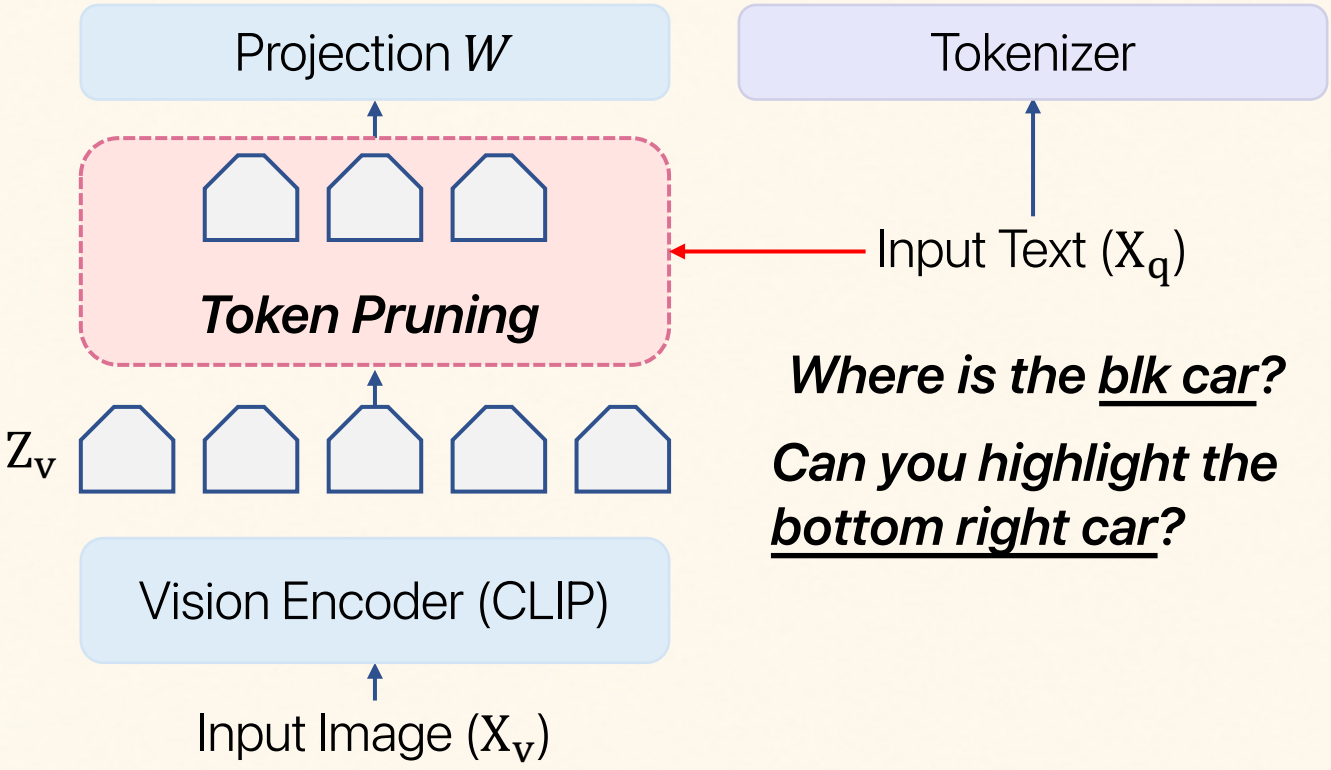
VisionZip[5]
(CVPR'25)



[5] Yang et al., VisionZip ···, CVPR, 2025.

The Answer Lies in Text

- As visual token importance varies with the input text, we need text-guided token pruning.



Revisiting CLIP

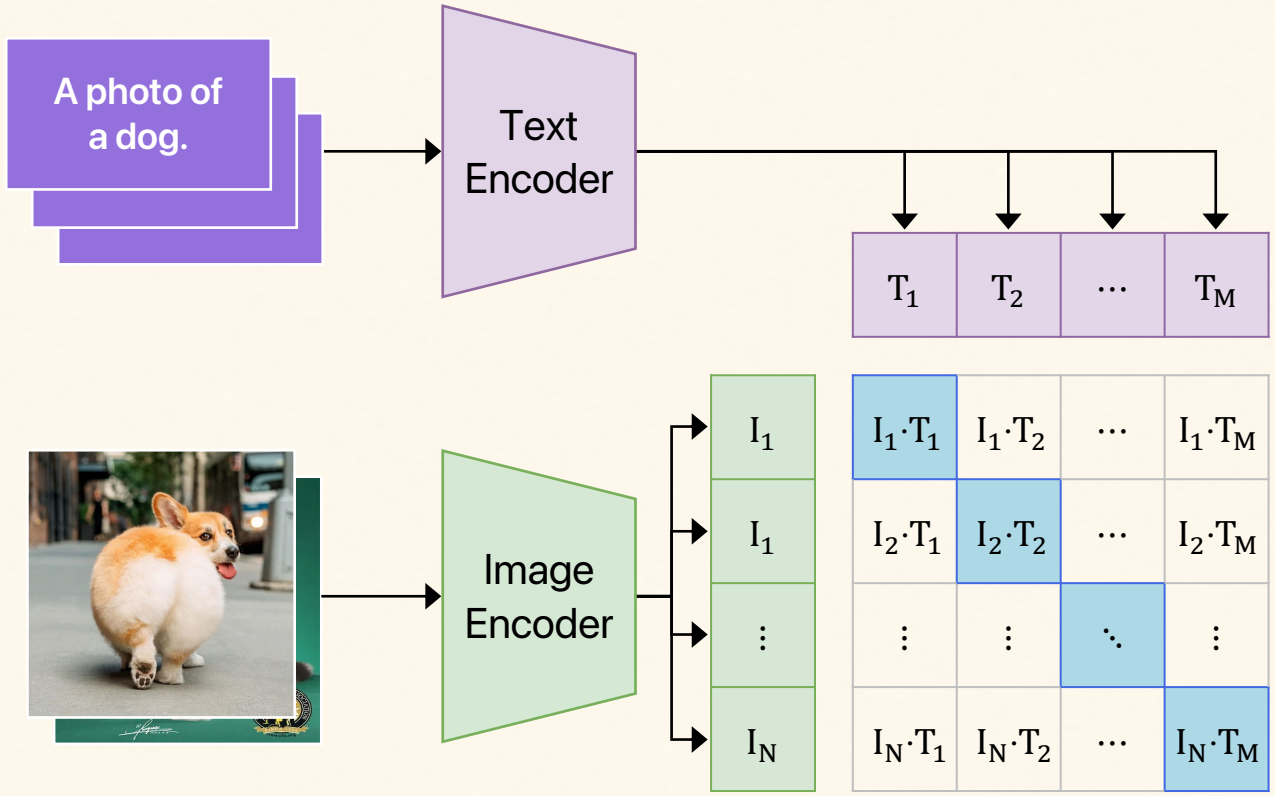
- To this end, we closely analyze CLIP.

Revisiting CLIP

- To this end, we closely analyze CLIP.
- CLIP is one of the most widely used vision encoders in LVLMs.

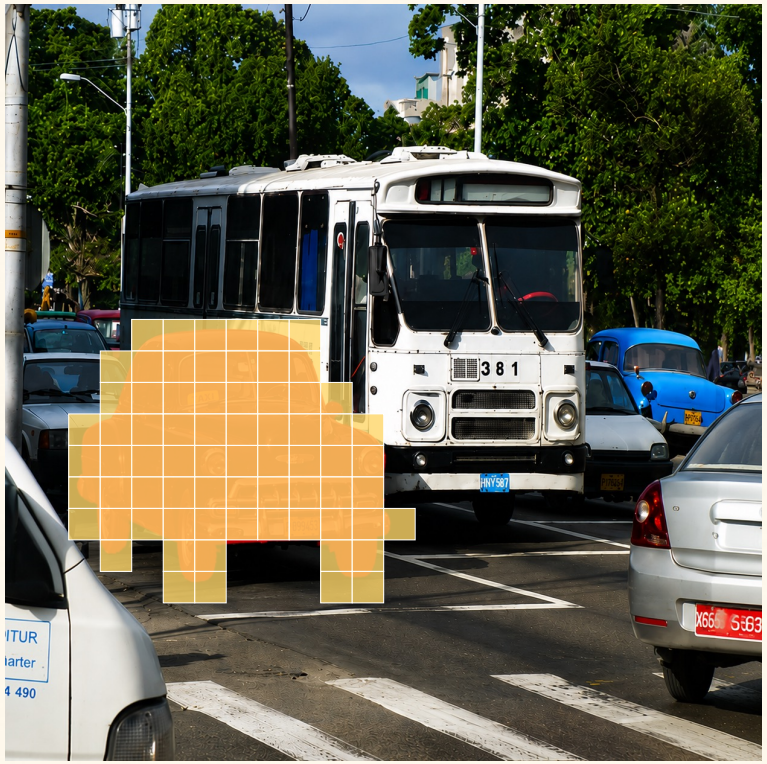
Revisiting CLIP

- To this end, we closely analyze CLIP.
- CLIP is one of the most widely used vision encoders in LVLMs.
- CLIP consists of a vision encoder and a text encoder.



CLIP Visual-Text Similarity

- Intuitively, visual tokens within the referent are expected to have higher similarity with the text.



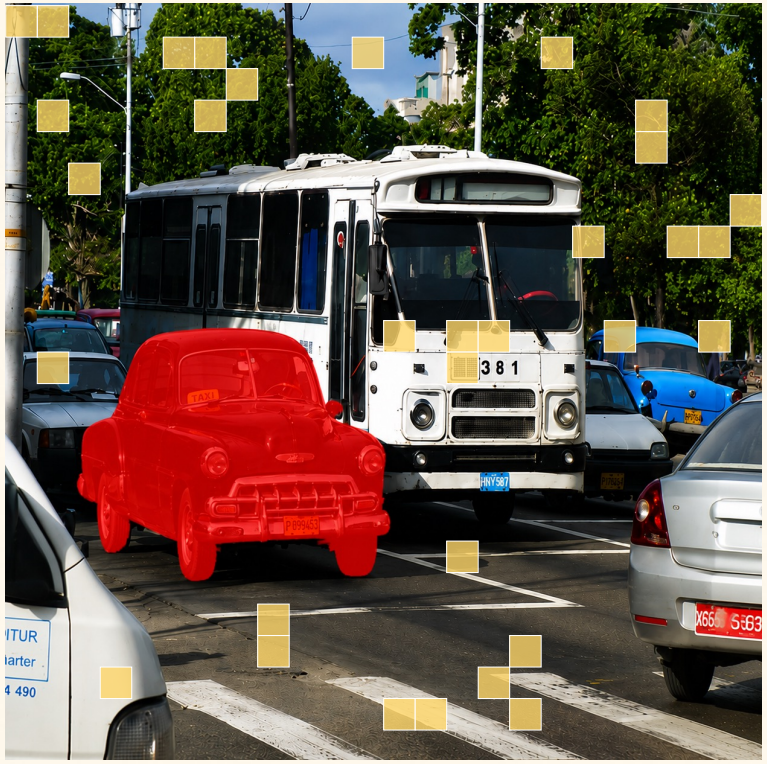
blk car

bottom right car

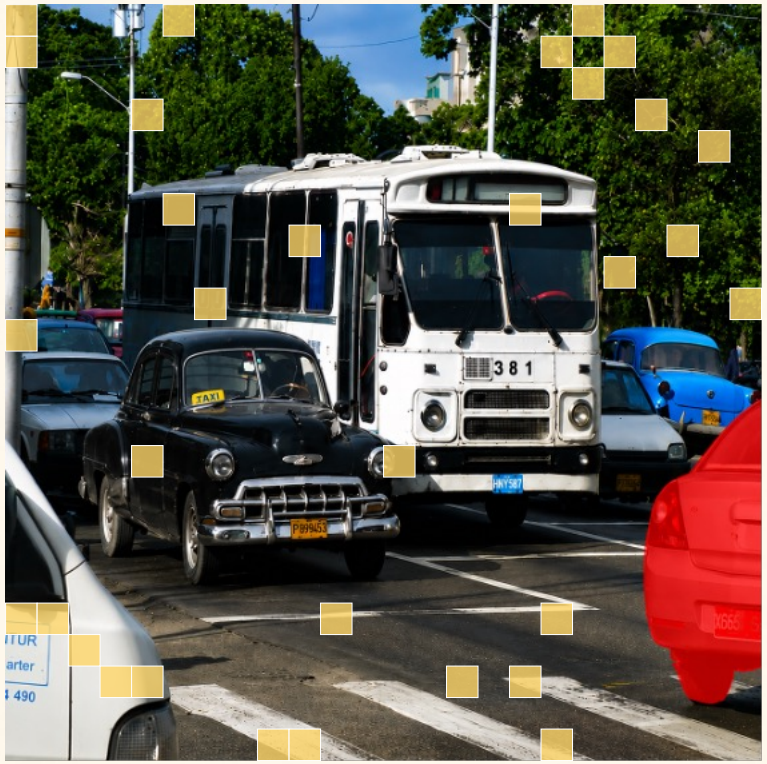
Visual-Text Similarity Reversal

- Intuitively, visual tokens within the referent are expected to have higher similarity with the text.
- However, we find that these tokens unexpectedly show low similarity.

Top-5% Visual-Text Similarity Tokens



blk car



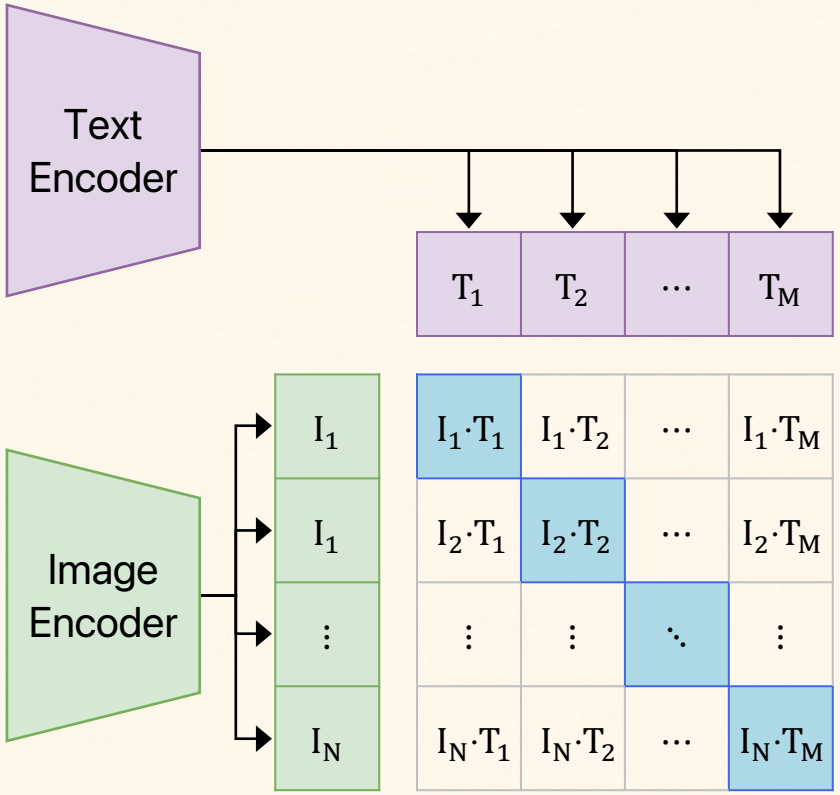
bottom right car

Visualization

Fig. Visualization of CLIP Visual-Text Similarity Reversal (RefCOCO+/g).

What Causes the Reversal?

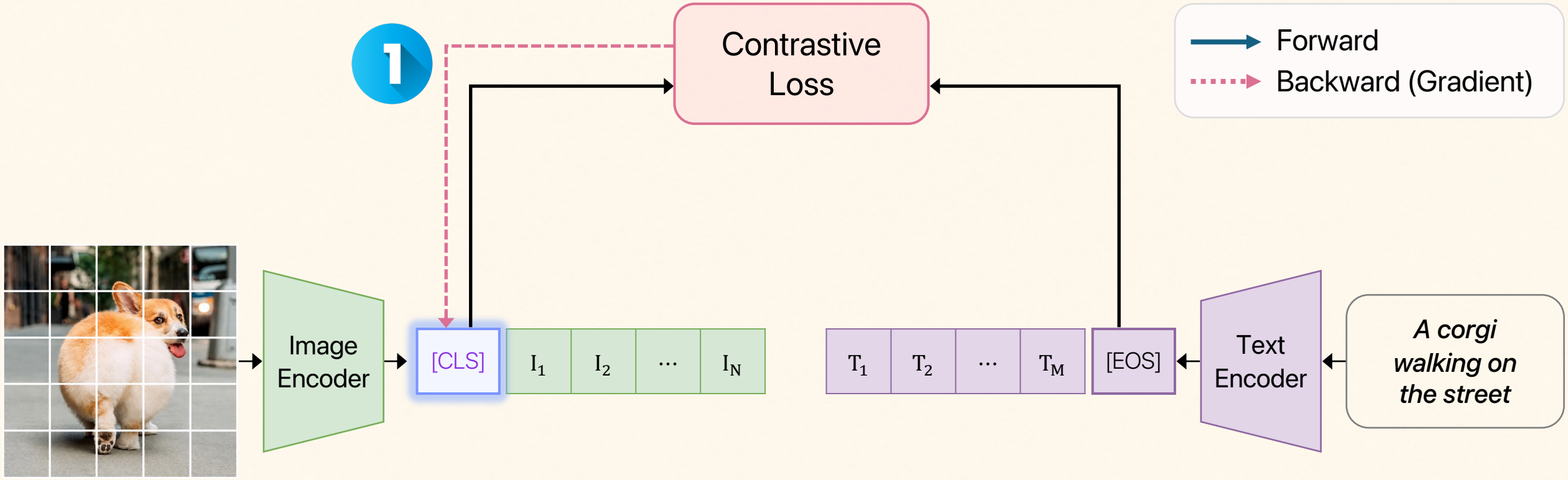
- CLIP is pretrained on millions of image-text pairs with contrastive learning.



Contrastive Learning

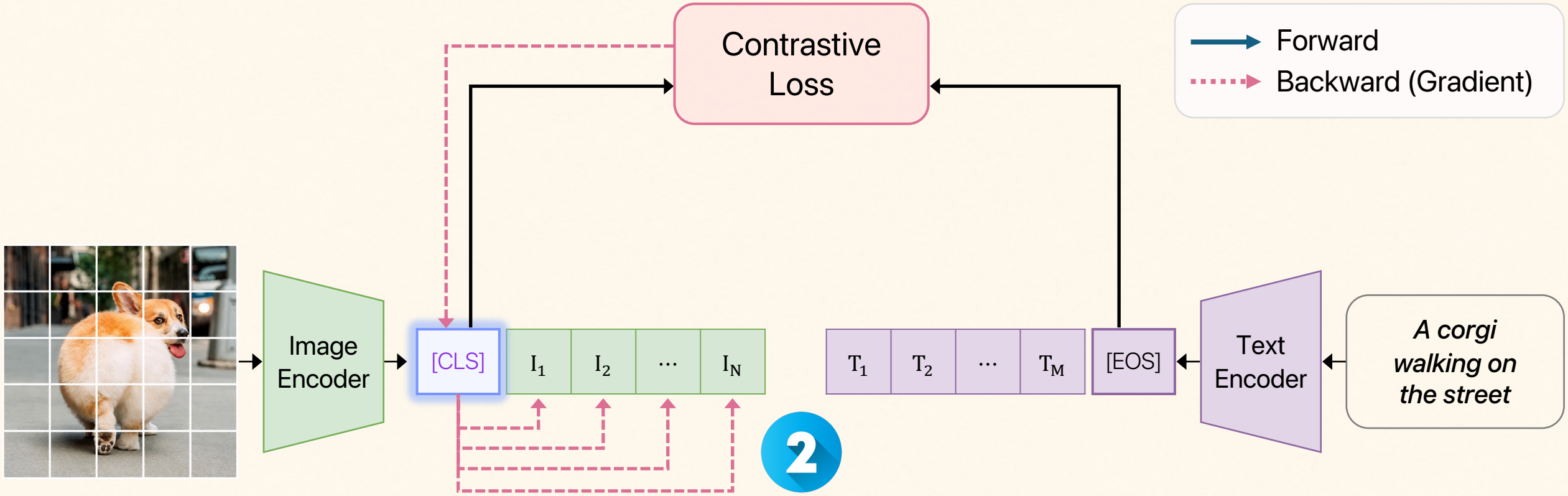
What Causes the Reversal?

➢ (1) During pretraining, image-to-text gradients flow into the [CLS] token.



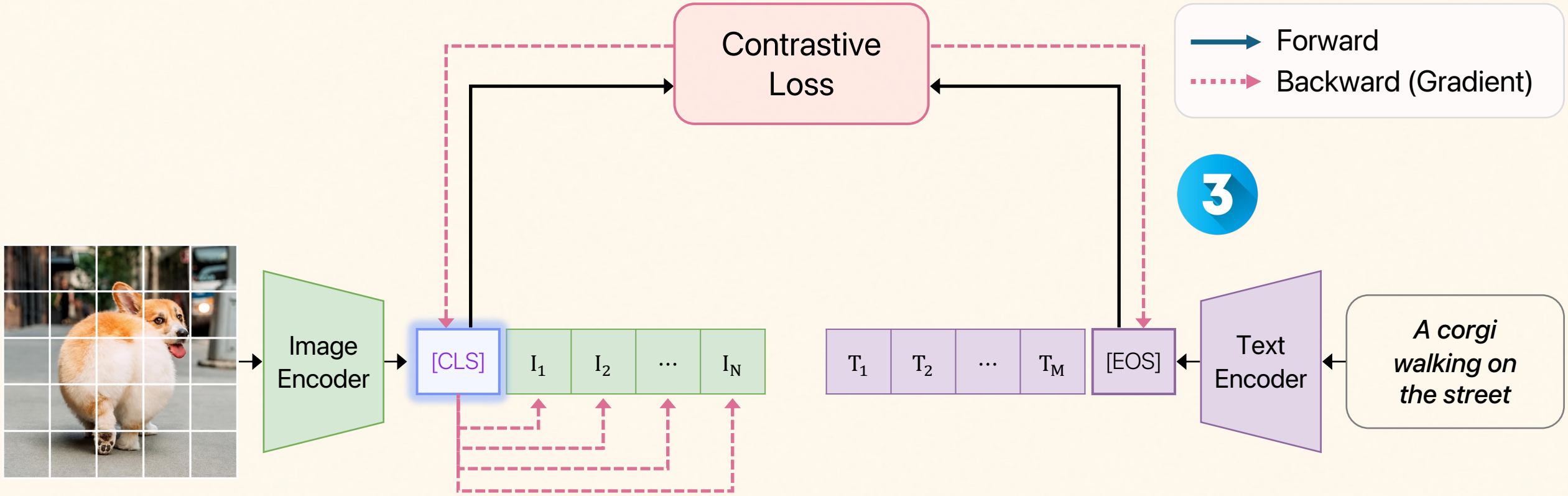
What Causes the Reversal?

- During pretraining, image-to-text gradients flow into the [CLS] token.
- (2) Following the chain rule, these gradients are further backpropagated to all visual tokens.



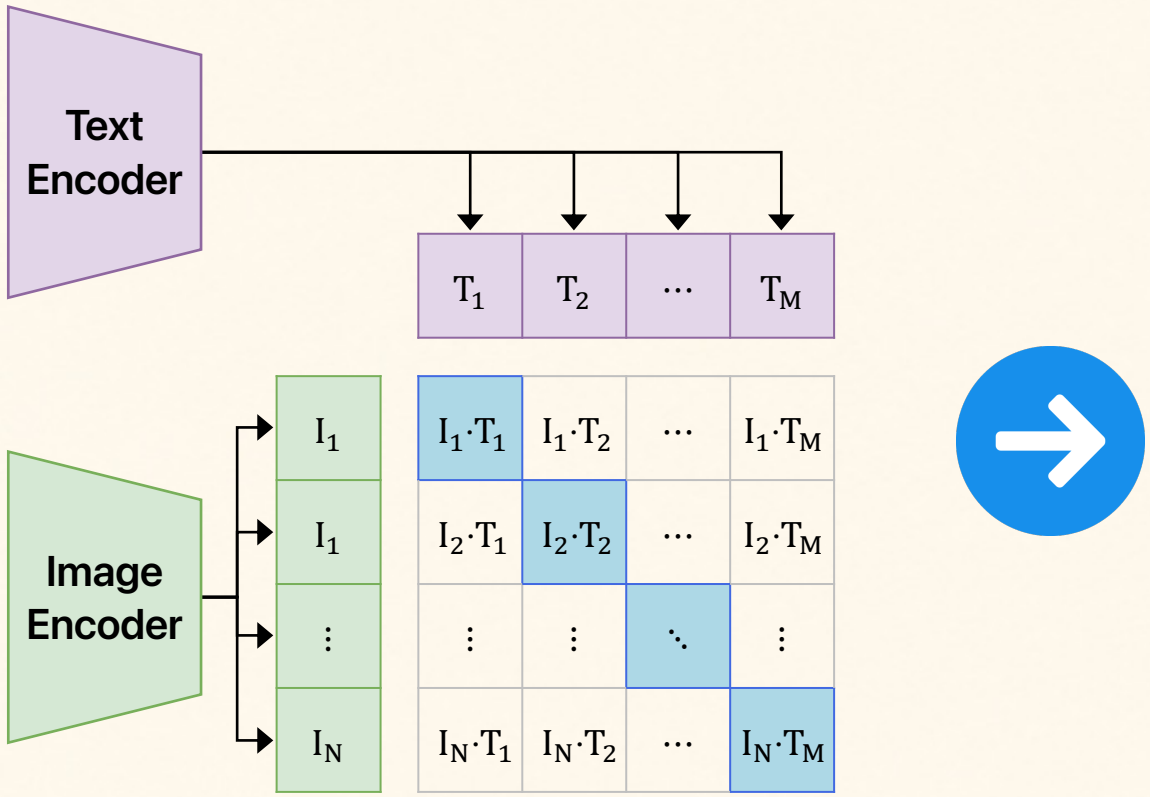
What Causes the Reversal?

- During pretraining, image-to-text gradients flow into the [CLS] token.
- Following the chain rule, these gradients are further backpropagated to all visual tokens.
- (3) Meanwhile, these gradients also flow into the [EOS] token.

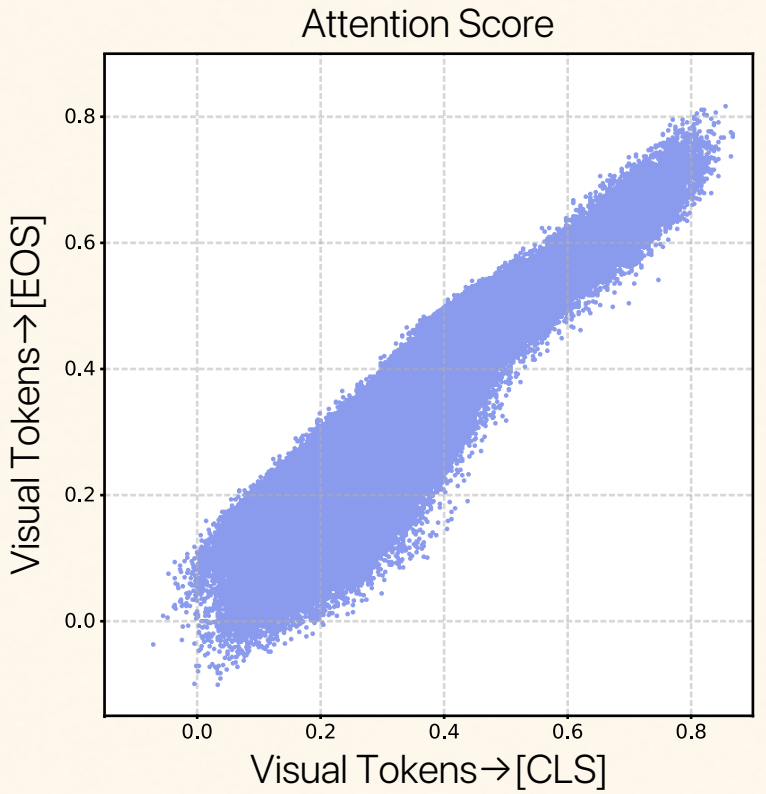


What Causes the Reversal?

- As a result, visual tokens with high [CLS] attention also show high [EOS] attention.



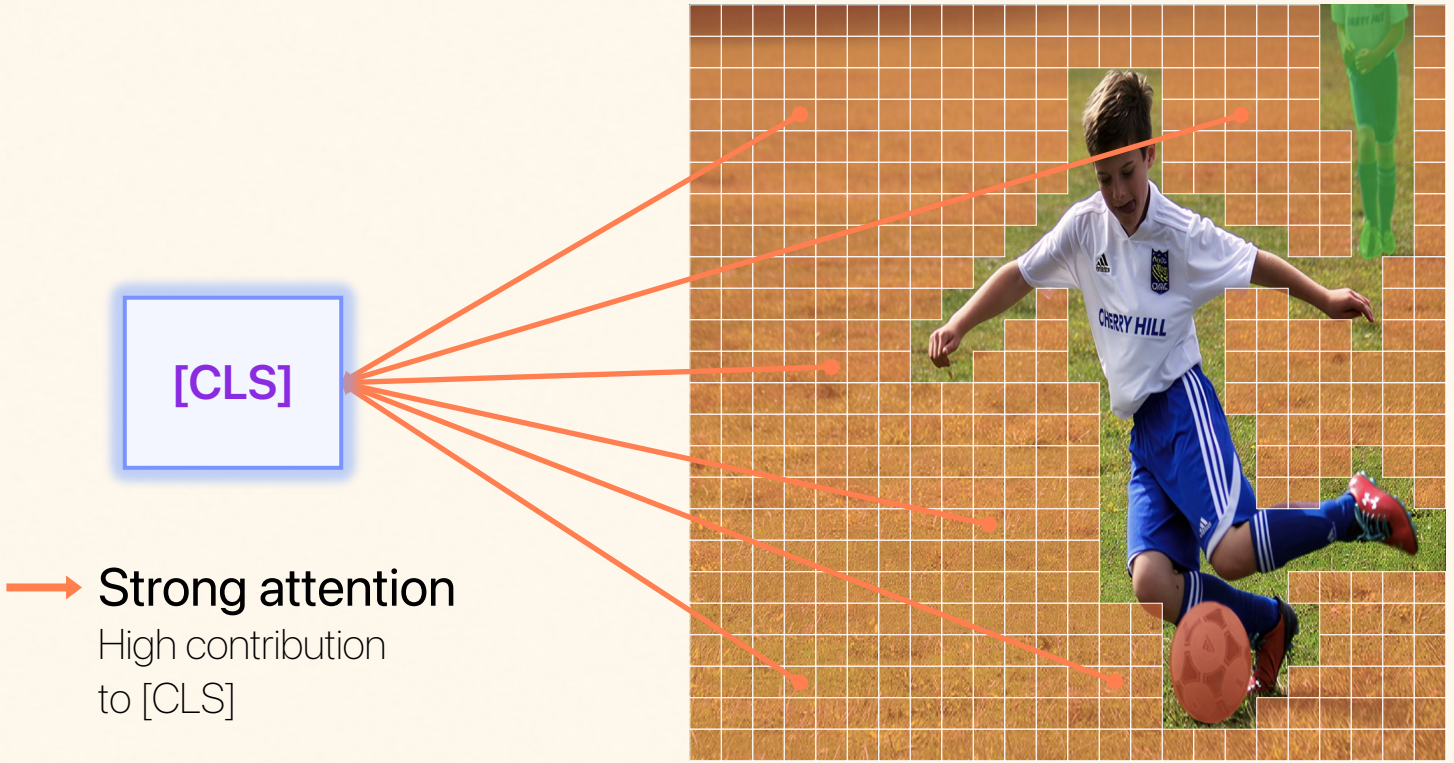
Contrastive Learning



[CLS]-[EOS] Attention Correlation

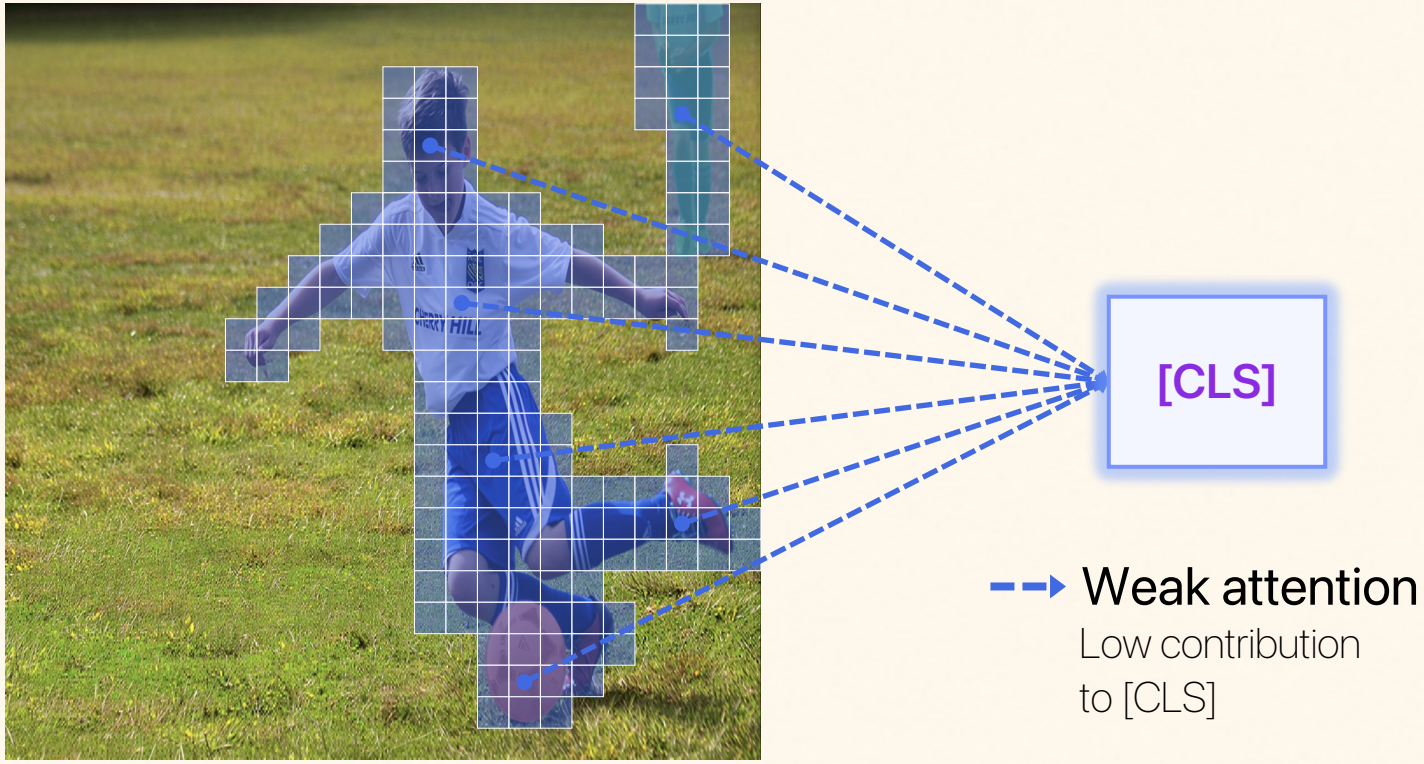
What Causes the Reversal?

- Background tokens strongly attend (similar) to [CLS].



What Causes the Reversal?

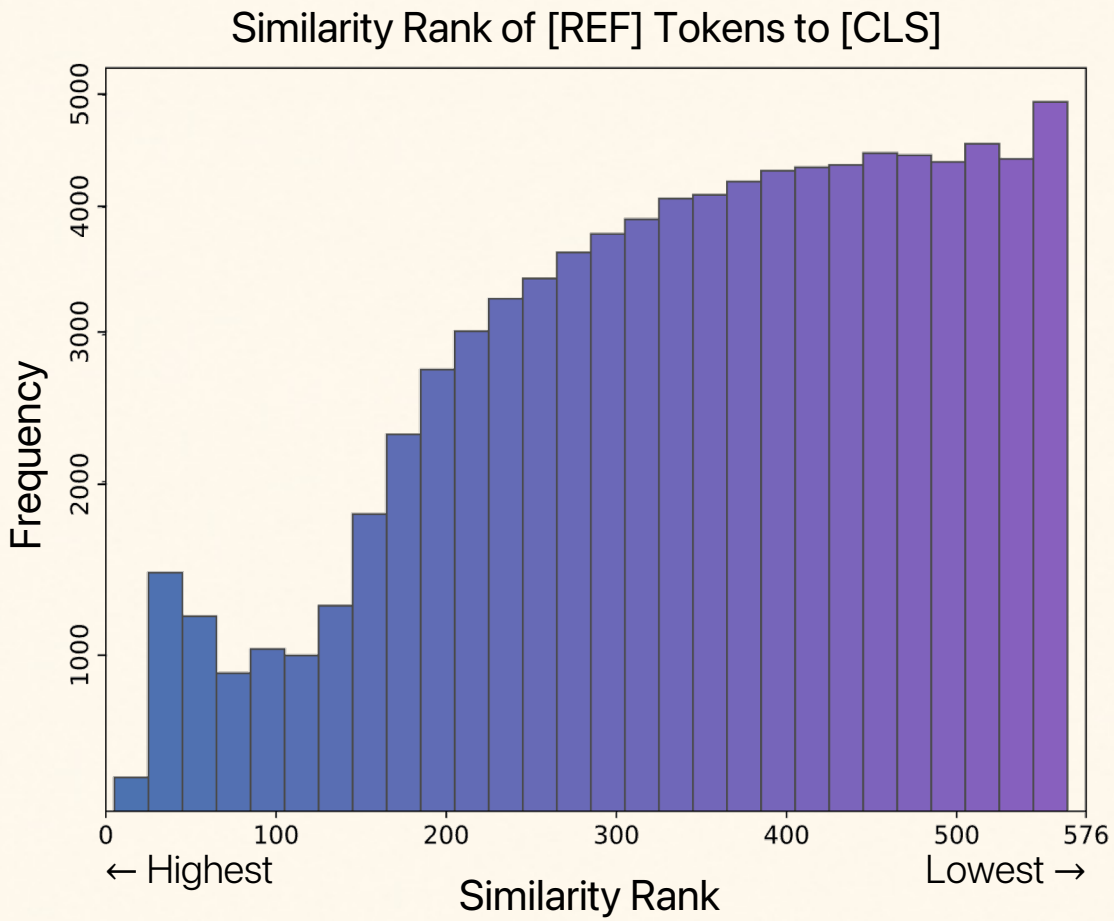
- Background tokens strongly attend (similar) to [CLS].
- Object tokens weakly attend (dissimilar) to [CLS].



Object Tokens
(e.g., player, soccer ball)

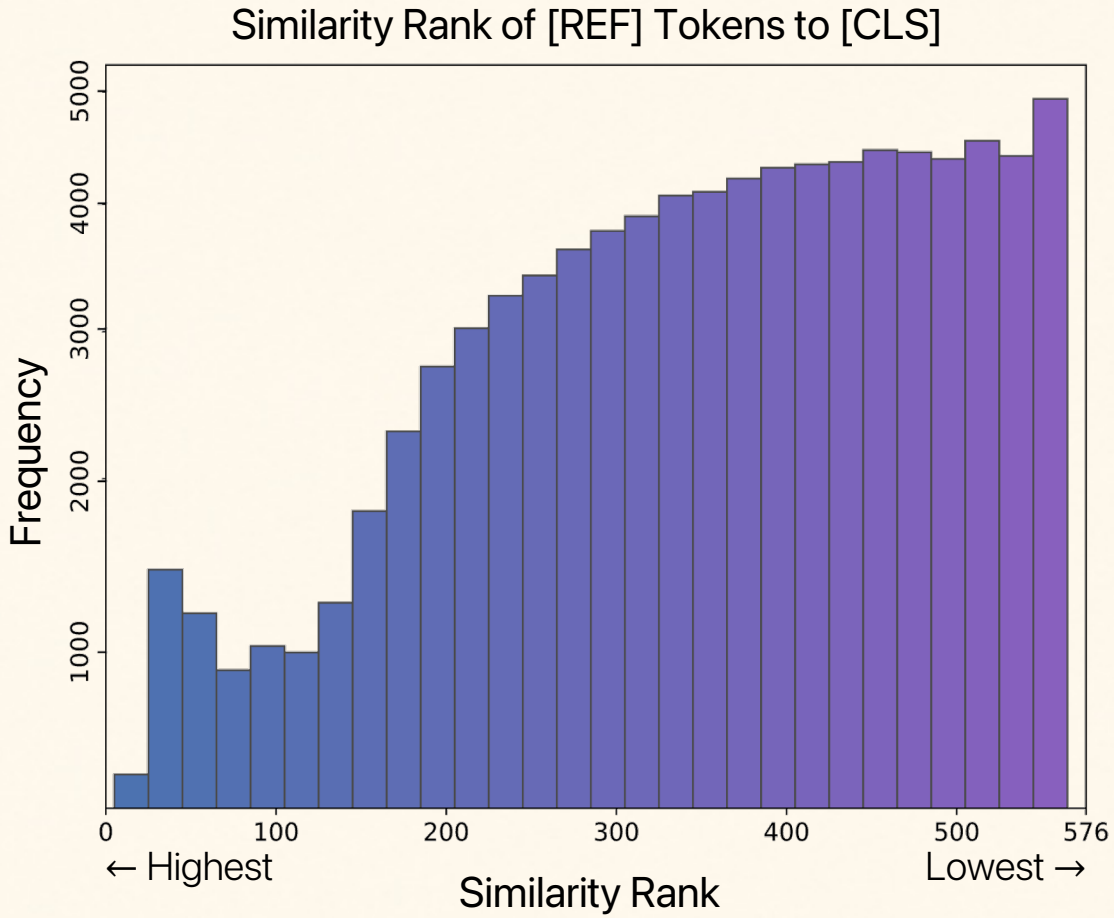
What Causes the Reversal?

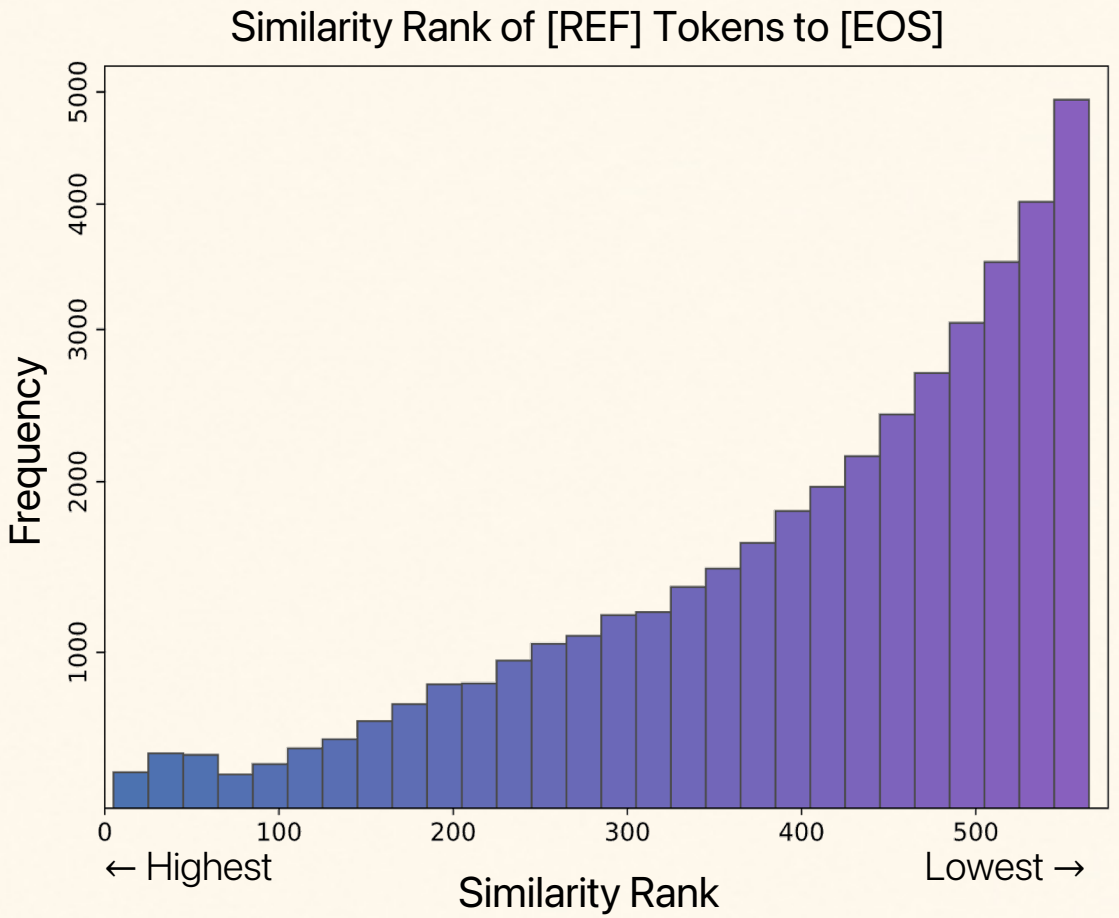
- [REF] tokens show low similarity to [CLS].



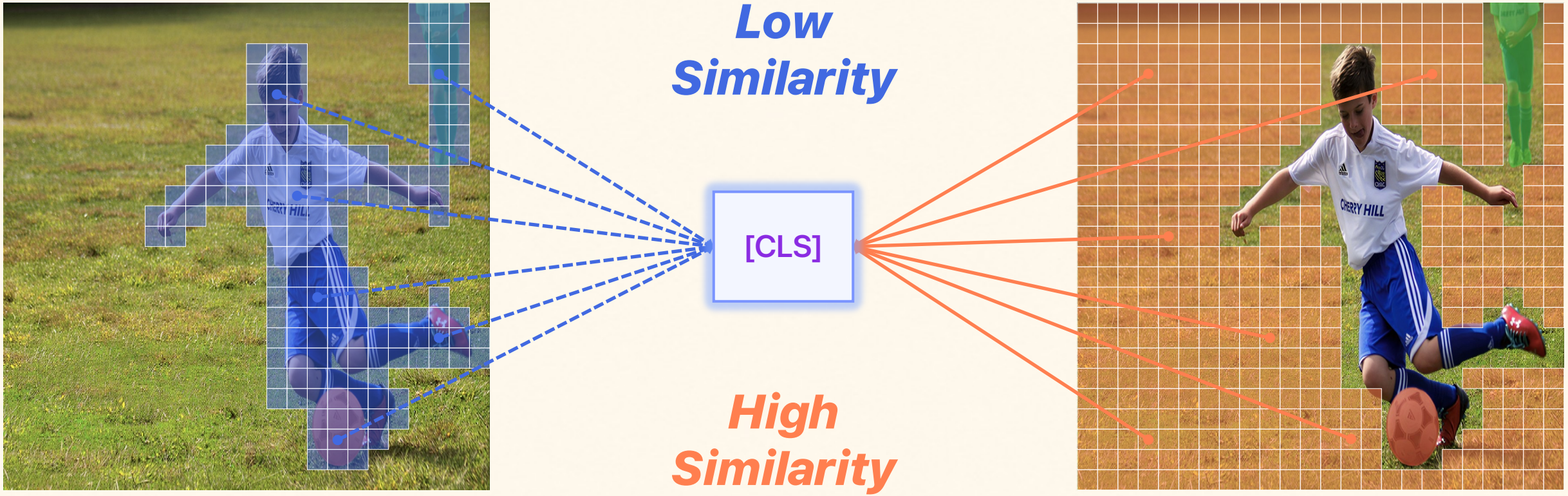
What Causes the Reversal?

- [REF] (object) tokens show low similarity to [CLS].
- [REF] tokens show even lower similarity to [EOS].





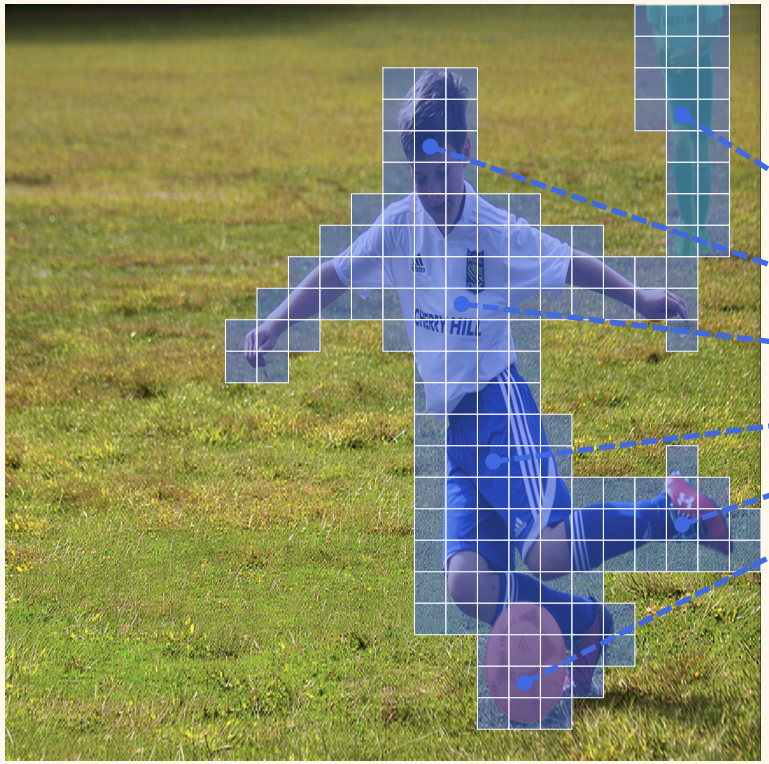
What Causes the Reversal?



Object Tokens
(e.g., grassy field)

Background Tokens
(e.g., player, soccer ball)

What Causes the Reversal?

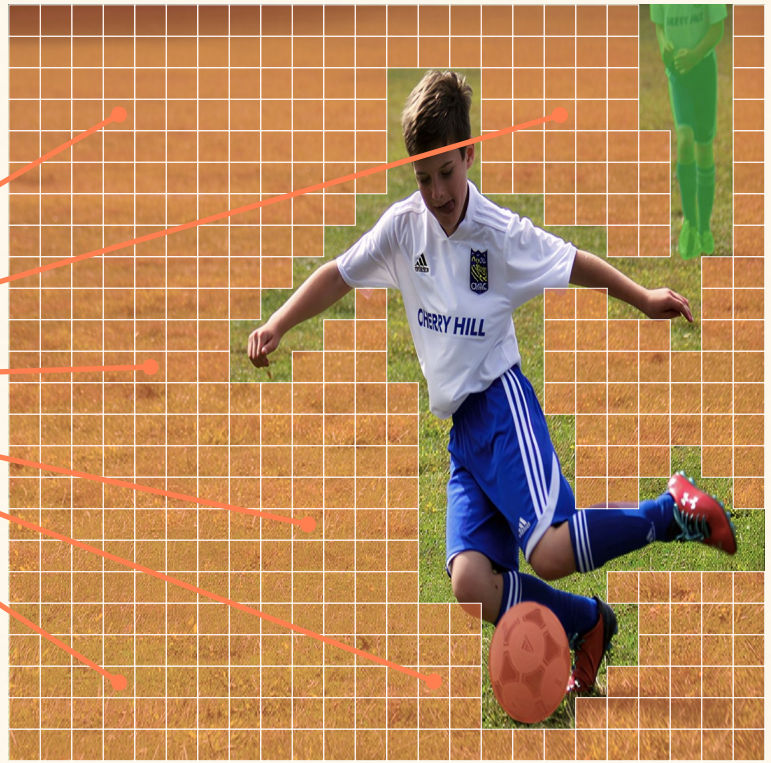


Object Tokens
(e.g., grassy field)

Even Lower Similarity

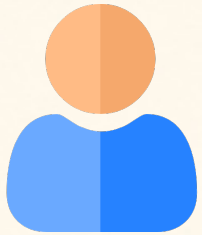
[EOS]

Even Higher Similarity



Background Tokens
(e.g., player, soccer ball)

What Accelerates the Reversal?



Please segment
the player wearing green cleats

CLIP
Text Encoder

What Accelerates the Reversal?

Please segment
the player wearing green cleats

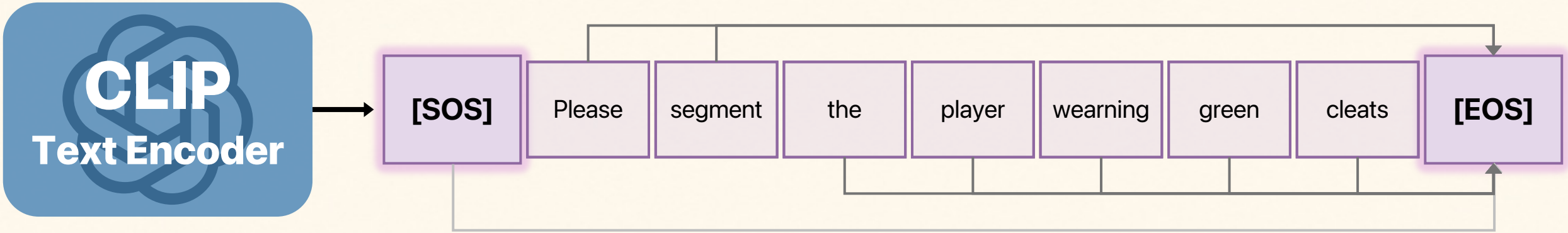
CLIP
Text Encoder

[SOS] | T_1 | T_2 | ... | T_M | [EOS]

please — *cleats* — *segment*

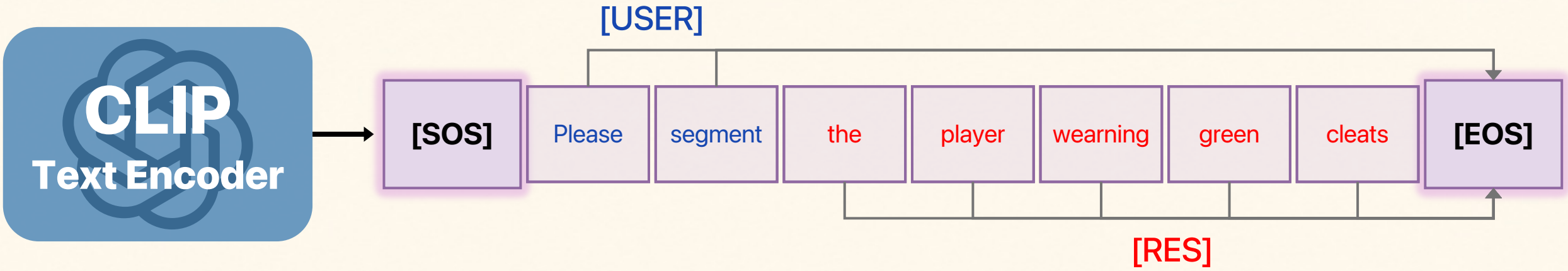
What Accelerates the Reversal?

- [EOS] token summarizes all text semantics into a single token.



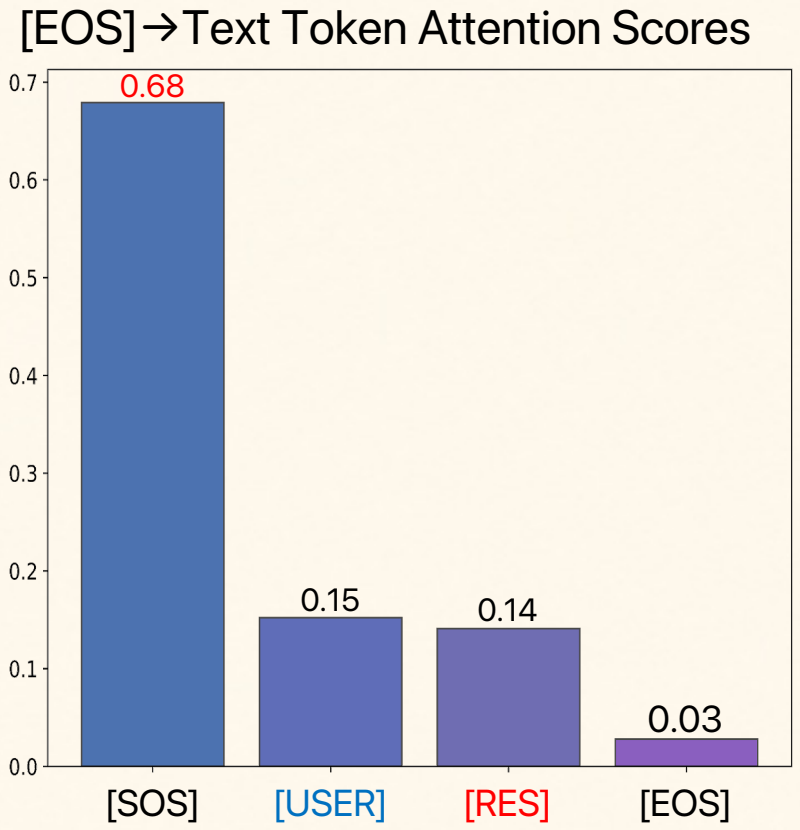
What Accelerates the Reversal?

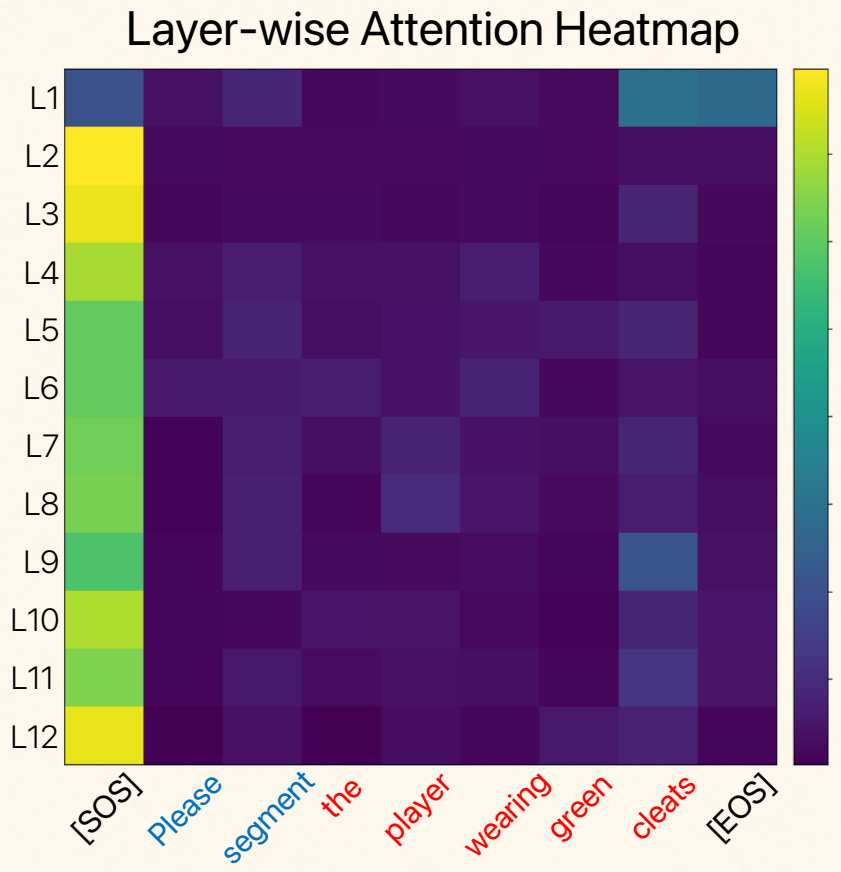
- [EOS] token summarizes all text semantics into a single token.
- Intuitively, [EOS] is expected to strongly attend to referent text tokens-[RES].



Text Attention Sink

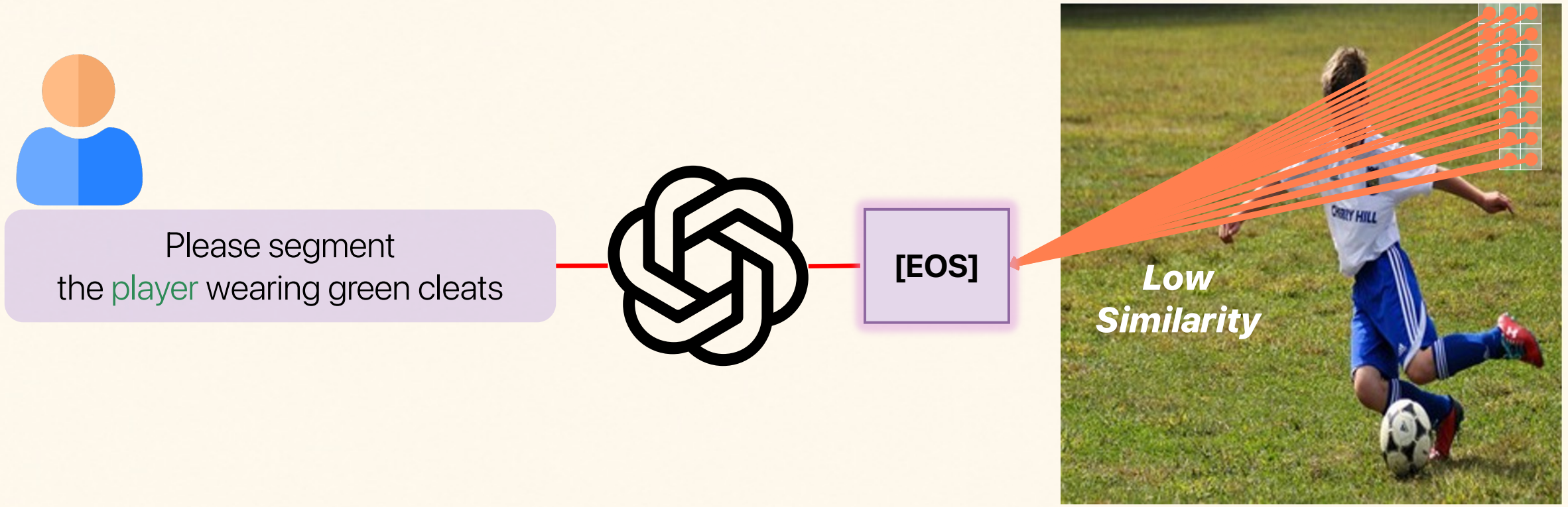
- [EOS] token summarizes all text semantics into a single token.
- Intuitively, [EOS] is expected to strongly attend to referent text tokens-[RES].
- However, we find [EOS] is heavily biased toward the [SOS] token.





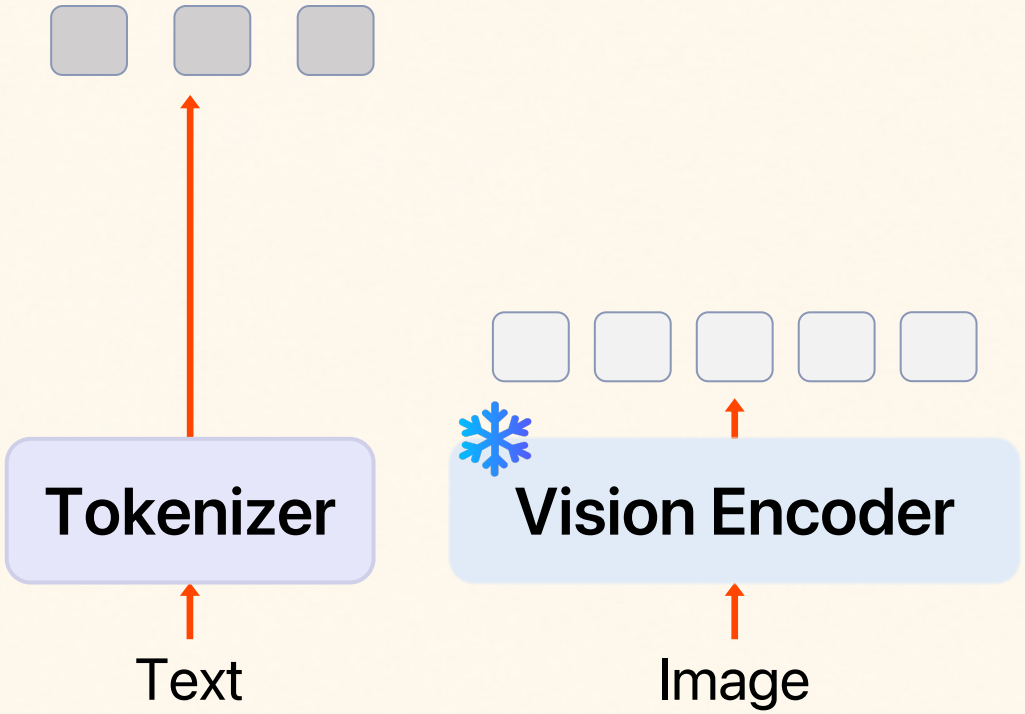
What Accelerates the Reversal?

- However, we find [EOS] is heavily biased toward the [SOS] token.
- Taken together, text attention sink accelerates the visual-text similarity reversal.



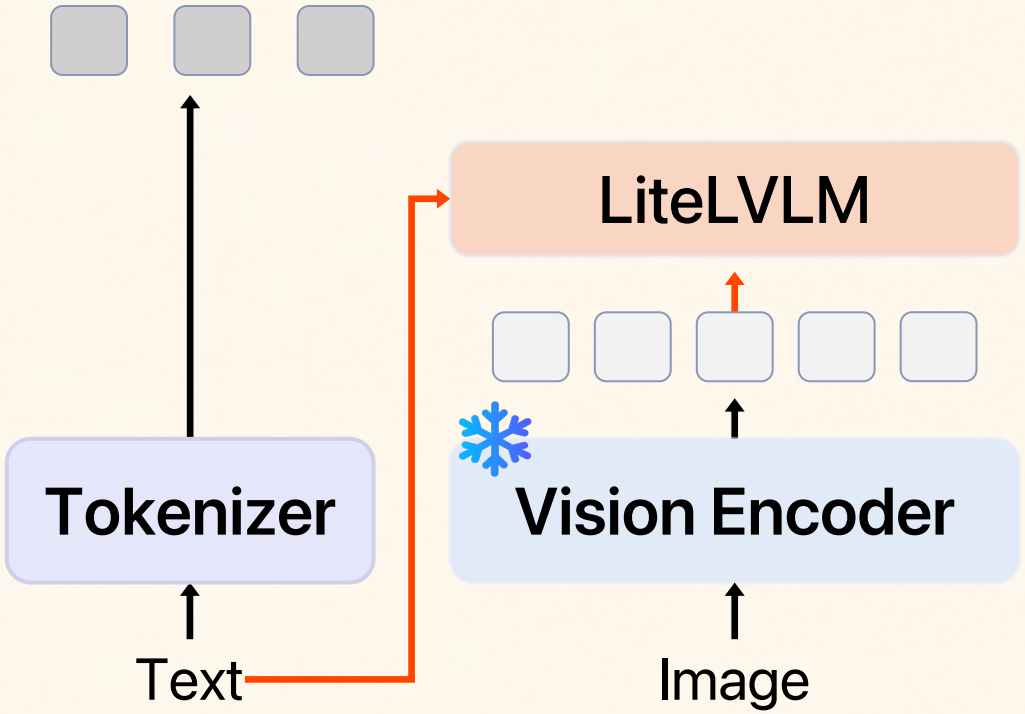
LiteLVLM

- Tokenizer and vision encoder encode the text and image into tokens.



LiteLVLM

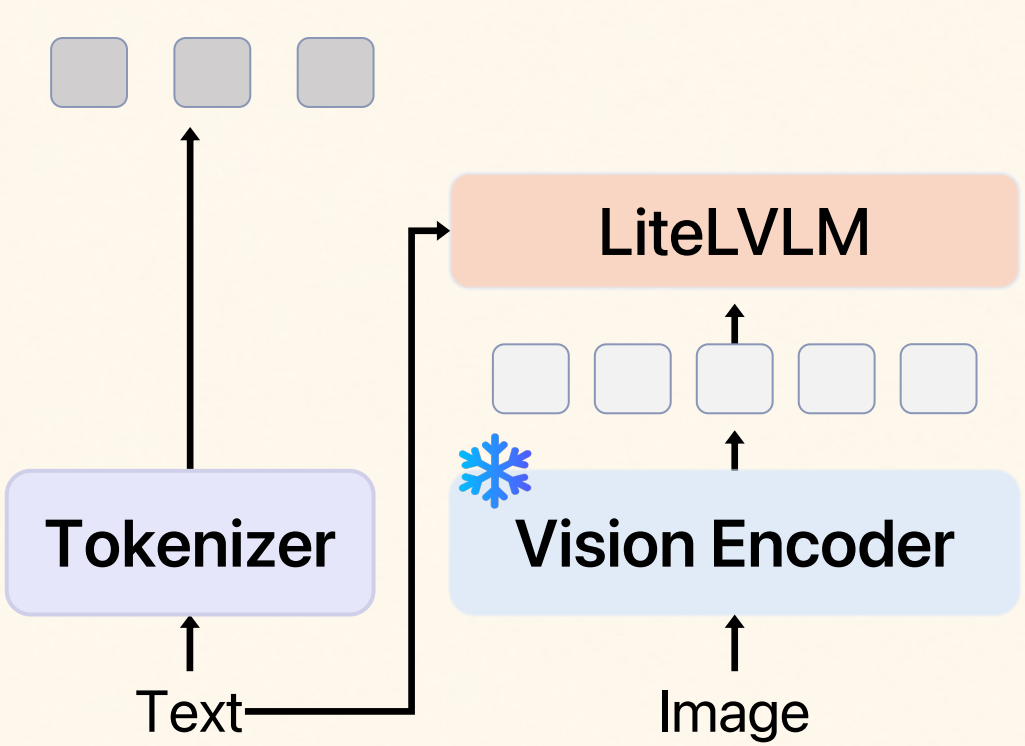
- LiteLVLM takes text and visual tokens.

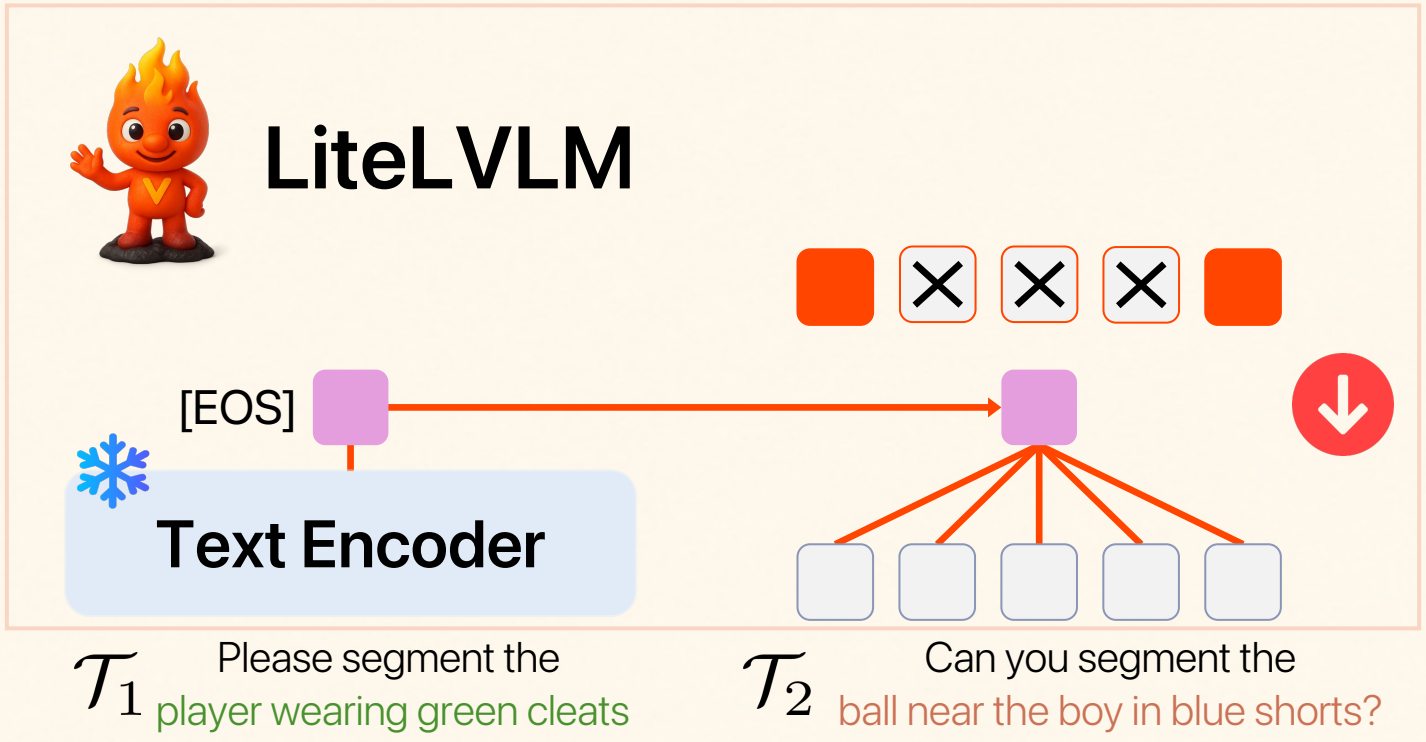


LiteLVLM: Similarity-aware Token Selection

- First, LiteLVLM select visual tokens with low [EOS] similarity.

$$s_{i,j} = E_i^{\mathcal{T}} \cdot E_j^{v\top}, \quad j = 1, \ldots, M.$$

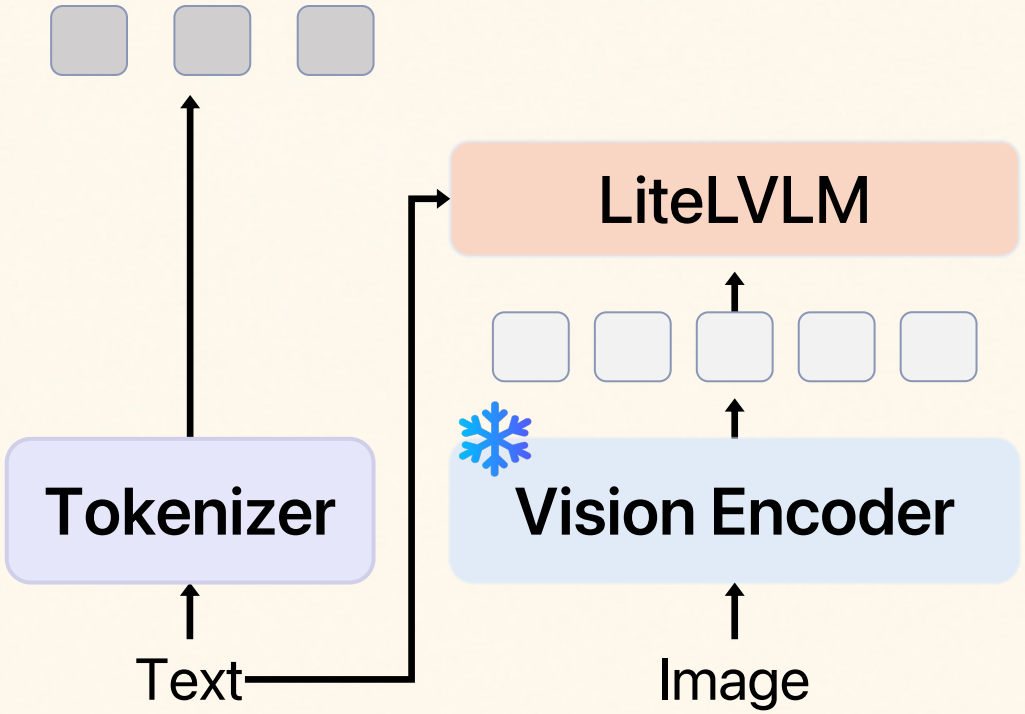


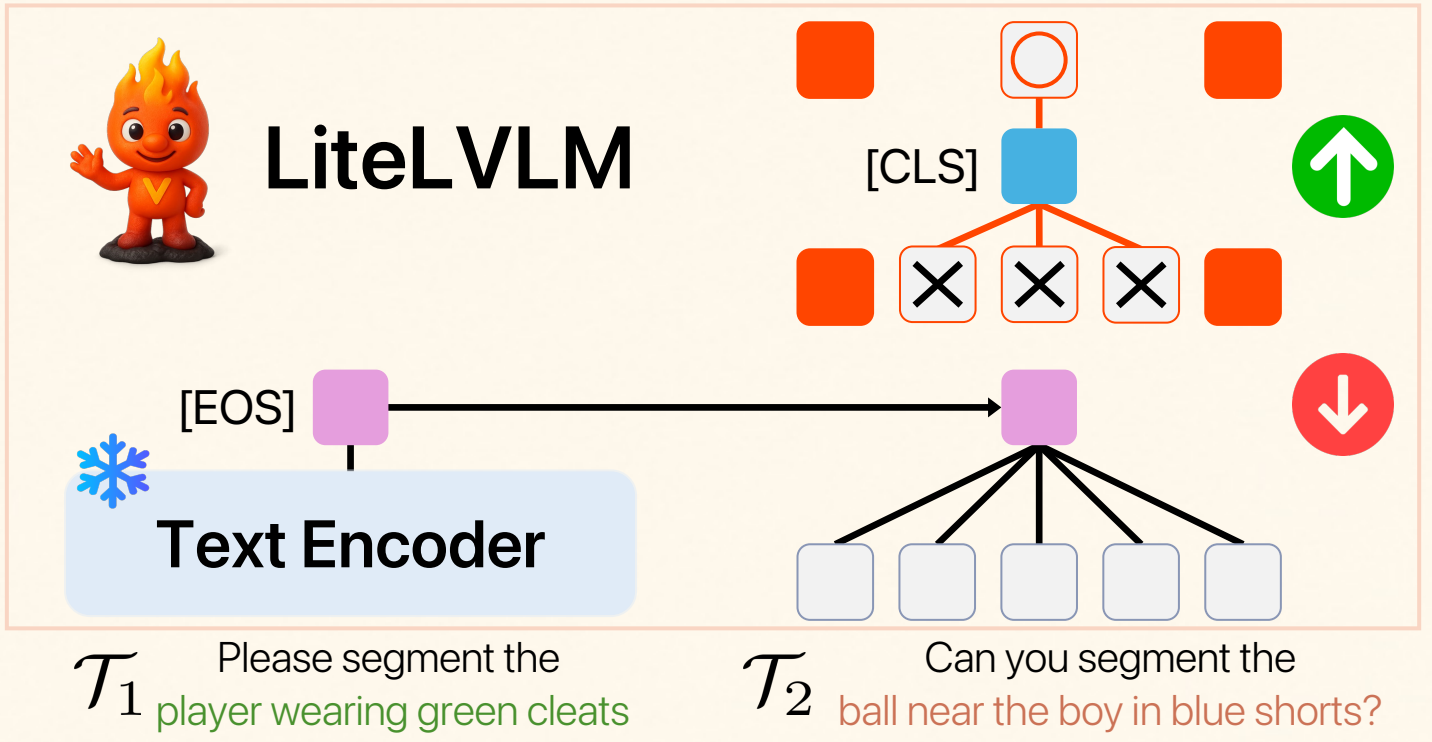


LiteLVLM: Context-aware Token Recovery

- Then, recover contextually informative tokens according to token contributions to [CLS].

$$s_i' = \left\| \frac{\exp(Q^{\mathcal{I}} K_i^{\top} / \sqrt{d})}{\sum_{j \in \mathcal{V}} \exp(Q^{\mathcal{I}} K_j^{\top} / \sqrt{d})} \cdot V_i \right\|_2$$





LiteLVLM

➢ The LLM and pixel decoder generate pixel-level mask from retained tokens.

$\mathcal{T}_1$ player

$\mathcal{T}_2$ ball

Language-to-Prompt (L-P)

Language Model (LLM)

Grounding Image Encoder

Pixel Decoder

LiteLVLM

Tokenizer

Vision Encoder

Text

Image

LiteLVLM

[CLS]

[EOS]

Text Encoder

$\mathcal{T}_1$ Please segment the player wearing green cleats

$\mathcal{T}_2$ Can you segment the ball near the boy in blue shorts?

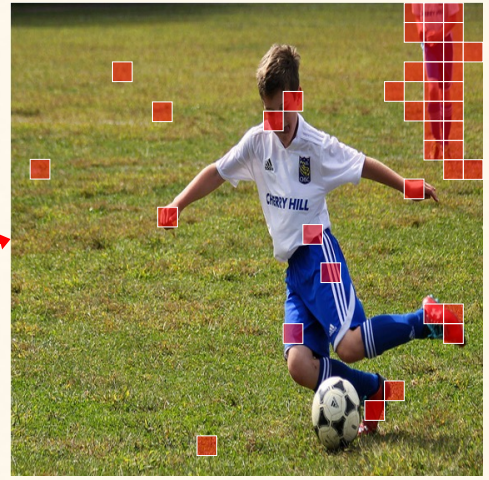
LiteLVLM: Adaptive Token Selection

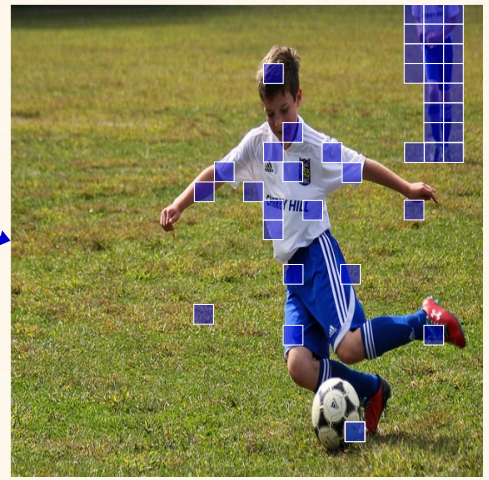
- For each input text, LiteLVLM first selects its corresponding token set.

$\mathcal{T}_1$ Please segment the player wearing green cleats

$\mathcal{T}_2$ Please highlight the distant player

LiteLVLM



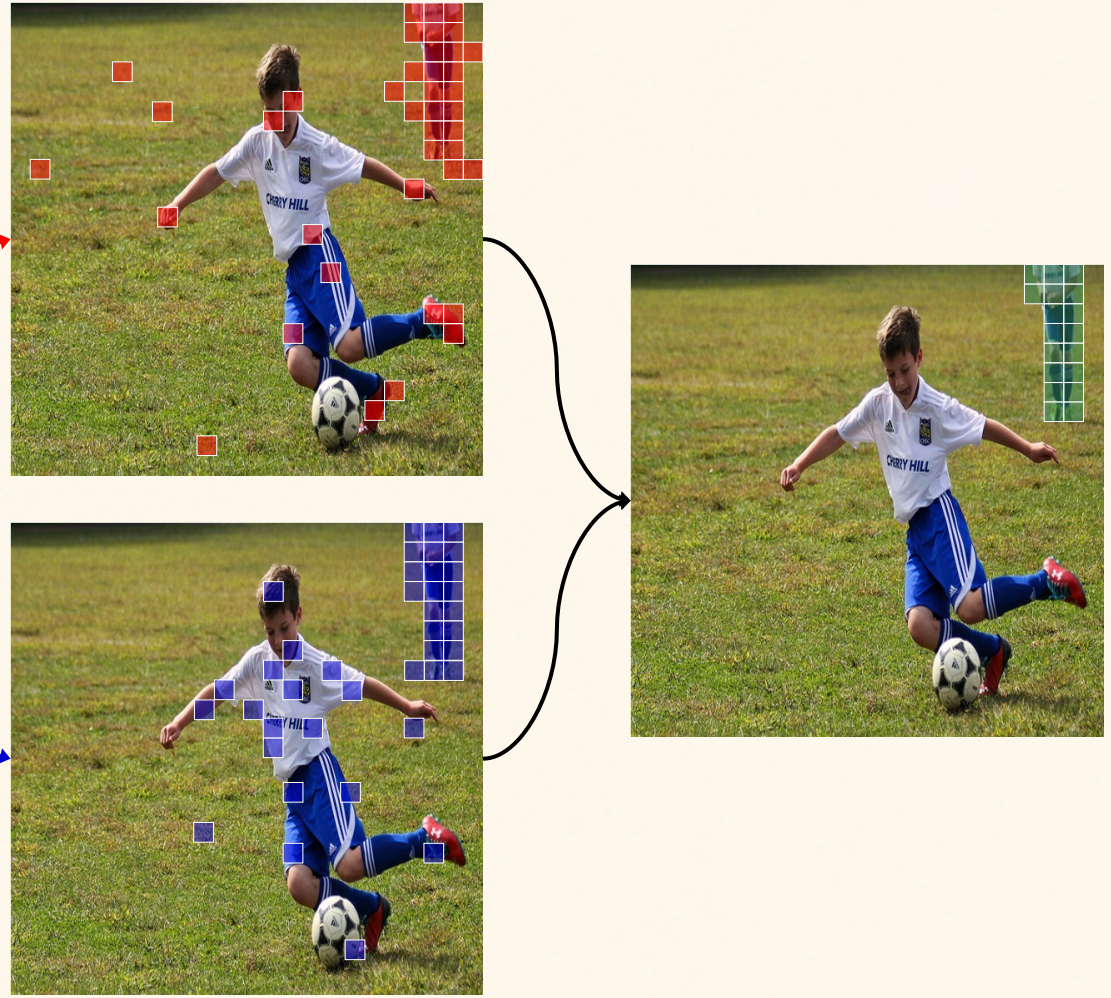


LiteLVLM: Adaptive Token Selection

- Then, we retain only the intersection to suppress noisy tokens.

$\mathcal{T}_1$ Please segment the player wearing green cleats

$\mathcal{T}_2$ Please highlight the distant player



Referring Expression Segmentation

Tab. Performance comparison of LiteLVLM on RefCOCO series.

Method	RefCOCO			RefCOCO+			RefCOCOg		Avg.	Rel.
	val	testA	testB	val	testA	testB	val	test		
*Upper Bound, All **576** Tokens* (**100%**)										
GLaMM[6]	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	75.5	100%
*Retain **192** Tokens* (↓ **66.7%**)										
LLaVA-PruMerge	68.8	74.6	64.1	58.2	66.2	50.0	61.2	62.1	63.1	83.5%
VisionZip	71.1	76.4	66.6	59.7	67.2	54.0	64.7	64.3	65.5	86.7%
LiteLVLM (ICML26)	**74.4**	**78.7**	**67.0**	**64.1**	**72.2**	**55.2**	**66.0**	**67.8**	**68.1**	**90.3%**
*Retain **128** Tokens* (↓ **77.8%**)										
LLaVA-PruMerge	64.2	70.9	60.3	54.1	61.0	47.1	57.6	58.8	59.2	78.4%
VisionZip	66.4	71.1	60.9	54.5	62.2	47.9	59.6	59.0	60.2	79.7%
LiteLVLM (ICML26)	**72.1**	**77.5**	**64.5**	**61.7**	**69.0**	**52.0**	**63.3**	**63.7**	**65.5**	**86.8%**
*Retain **64** Tokens* (↓ **88.9%**)										
LLaVA-PruMerge	58.9	64.3	54.9	45.9	50.3	42.4	49.5	50.0	52.0	68.8%
VisionZip	59.0	63.8	55.1	47.1	51.9	40.1	49.2	51.7	52.2	69.1%
LiteLVLM (ICML26)	**66.3**	**74.5**	**58.2**	**56.2**	**64.0**	**46.7**	**56.1**	**56.5**	**59.8**	**79.2%**

[6] Rasheed et al., GLaMM: Pixel Grounding Large Multimodal Model, CVPR, 2024.

Referring Video Object Segmentation

Tab. Performance comparison on Ref-DAVIS-17.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Rel.
*Upper Bound, All **576** Tokens* (**100%**)				
VideoGLaMM[7]	65.6	73.3	69.5	100%
*Retain **196** Tokens* (↓ **65.9%**)				
VisPruner[8]	61.2	67.4	64.3	92.5%
LiteLVLM (ICML26)	**66.8**	**71.6**	**69.2**	**99.5%**
*Retain **81** Tokens* (↓ **85.9%**)				
VisPruner	57.1	63.5	60.3	86.7%
LiteLVLM (ICML26)	**64.3**	**67.8**	**66.1**	**95.1%**

Tab. Performance comparison on Refer-YouTube-VOS.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Rel.
*Upper Bound, All **576** Tokens* (**100%**)				
VideoGLaMM	65.4	68.2	66.8	100%
*Retain **196** Tokens* (↓ **65.9%**)				
VisPruner	60.6	63.9	62.3	93.2%
LiteLVLM (ICML26)	**65.1**	**67.9**	**66.5**	**99.5%**
*Retain **81** Tokens* (↓ **85.9%**)				
VisPruner	58.1	62.2	60.1	89.9%
LiteLVLM (ICML26)	**60.8**	**67.6**	**64.2**	**96.1%**

[7] Munasinghe et al., VideoGLaMM ···, CVPR, 2025.
[8] Zhang et al., Beyond Text-Visual Attention ···, ICCV, 2025.

Visualization

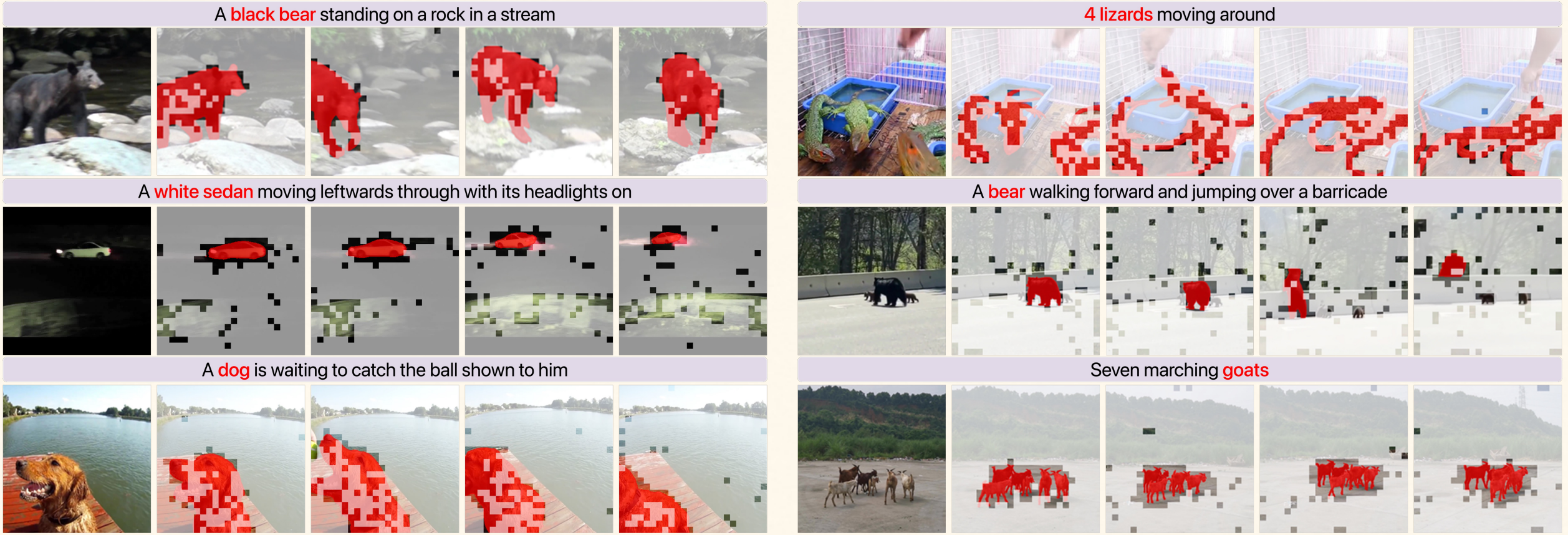


Fig. Visualizations of LiteLVLM on Refer-YouTube-VOS and MeViS.

How Efficient is LiteLVLM?

Tab. Performance comparison on Ref-DAVIS-17.

Method	FLOPs (TB)	Prefilling Time (ms)	CUDA Time (ms)	Storing Activation (GB)
*Upper Bound, All **576** Tokens* (**100%**)				
GLaMM	4.66	166.25	340.89	0.81
*Retain **192** Tokens* (↓ **66.7%**)				
FastV[9]	2.65	162.88	340.25	0.81
VisPruner	2.17	75.65	276.09	0.37
LiteLVLM (ICML26)	**2.11**	**74.88**	**265.83**	**0.35**
*Retain **64** Tokens* (↓ **88.9%**)				
FastV	2.23	157.50	338.65	0.80
VisPruner	1.33	54.77	253.10	0.23
LiteLVLM (ICML26)	**1.27**	**54.02**	**237.35**	**0.21**

[9] Chen et al., An Image is Worth 1/2 Tokens After Layer 2 ···, ECCV, 2024.

Thank You for Listening

Come Join Us!

paper

code

me