Exploring Unconventional Uses of LLMs in Vision Tasks

Anna Kukleva
Max-Planck-Institute for Informatics

The 50th Pattern Recognition and Computer Vision Colloquium

09.10.2025

LLMs are Everywhere

LLMs

Chatbots

Code Generation

Knowledge Retrieval

Summarization

LLMs are Everywhere

LLM driven Vision

LLMs are Everywhere

LLM driven Vision

- Object-based control in the real world [1]
- SMPL pose generation/editing [2]
- Tracking [3] and segmentation [4]
- Reasoning [5]
- …

[1] Learning to Generate Object Interactions with Physics-Guided Video Diffusion, Romero et al., arxiv
[2] UniPose: A Unified Multimodal Framework for Human Pose Comprehension, Generation and Editing, Li et al., CVPR 25
[3] Monocular-Video Based 3D Visual Language Tracking, Wei et. al, CVPR 25
[4] Unifying LLM-Driven Semantic Cues with Visual Features for Robust Few-Shot Segmentation, Karimi et al., CVPR 25
[5] Vision-Centric Reasoning with Grounded Chain-of-Thought, Man et al., CVPR 25

LLMs in this talk

LLM driven Vision

Fusion of LLM and:

- Diffusion models [1]
- Self-supervised vision pretraining [2]
- Large-scale video data [3]

[1] RefAM: Attention Magnets for Zero-Shot Referral Segmentation, Kukleva* & Simsar* et al., arxiv
[2] Language-Unlocked ViT (LUViT): Empowering Self-Supervised Vision Transformers with LLMs, Kuzucu et al., arxiv
[3] HowToCaption: Prompting LLMs to Transform Video Annotations at Scale, Shvetsova* & Kukleva* et al., ECCV 24

RefAM: Attention Magnets for Zero-Shot Referral Segmentation

Anna Kukleva[1*], Enis Simsar[2*], Alessio Tonioni[3], Ferjad Naeem[3], Federico Tombari[3,4] , Jan Eric Lenssen[1], Bernt Schiele[1]

[1]Max Planck Institute for Informatics, [2]ETH Zurich, [3]Google, [4]TU Munich

Leveraging pre-trained LLM

for implicit semantic understanding

Zero-Shot Referral Segmentation

A largest orange goldfish

Goal: given image/video and referral expression, segment corresponding objects in the image/video

Zero-Shot Referral Segmentation Pipeline

Previous work [1,2]

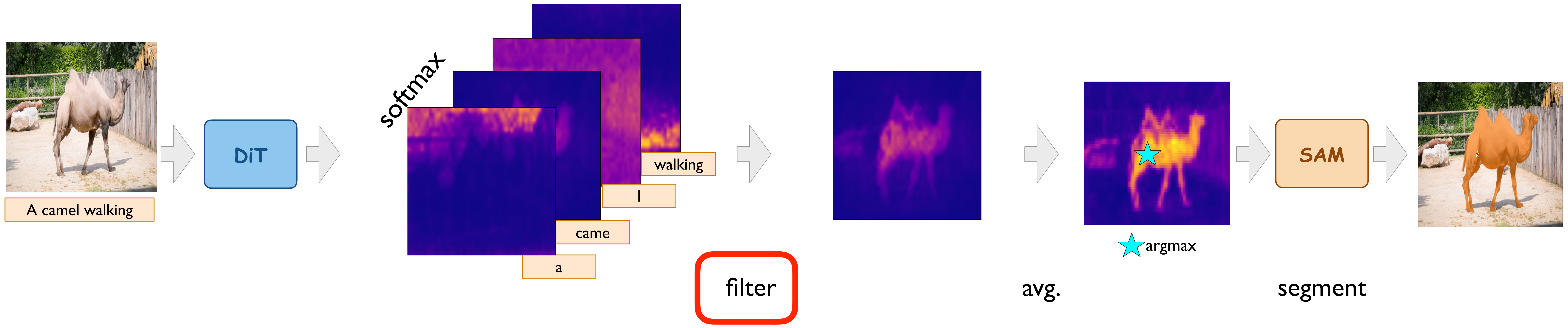
1. Mask proposals
2. Local and Global reasoning modules
3. Integration with CLIP visual-language space

Our work

1. **No** Mask proposal
2. **No** Local and Global reasoning modules
3. **No** Integration with CLIP visual-language space

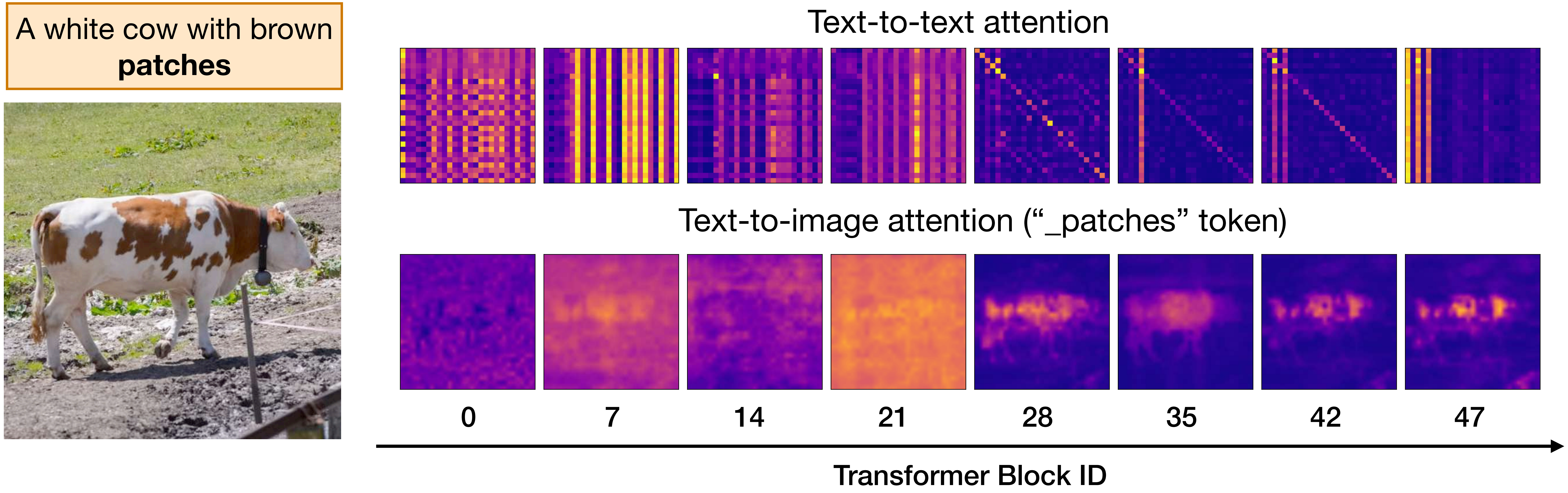
[1] Zero-Shot Referring Image Segmentation with Global-Local Context Features, Yu et al., CVPR 2023
[2] Hybrid Global-Local Representation with Augmented Spatial Guidance for Zero-Shot Referring Image Segmentation, Liu et al., CVPR 2025

Zero-Shot Referral Segmentation Pipeline

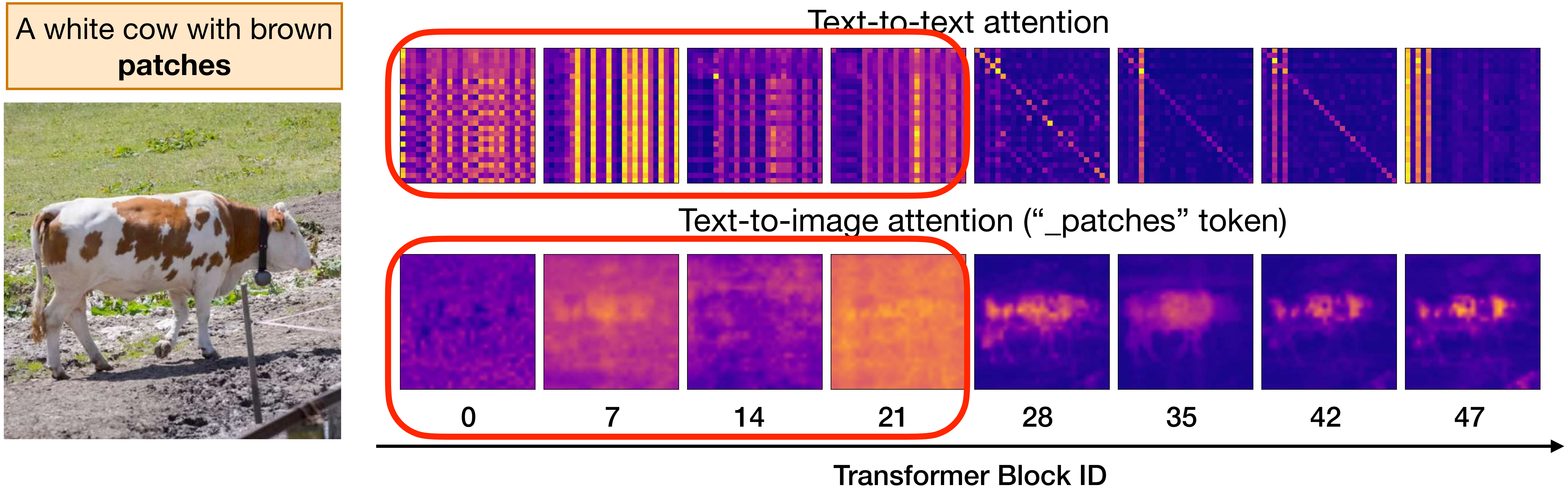


What and how we filter attention maps?

Emergence of Semantic Information in DiT

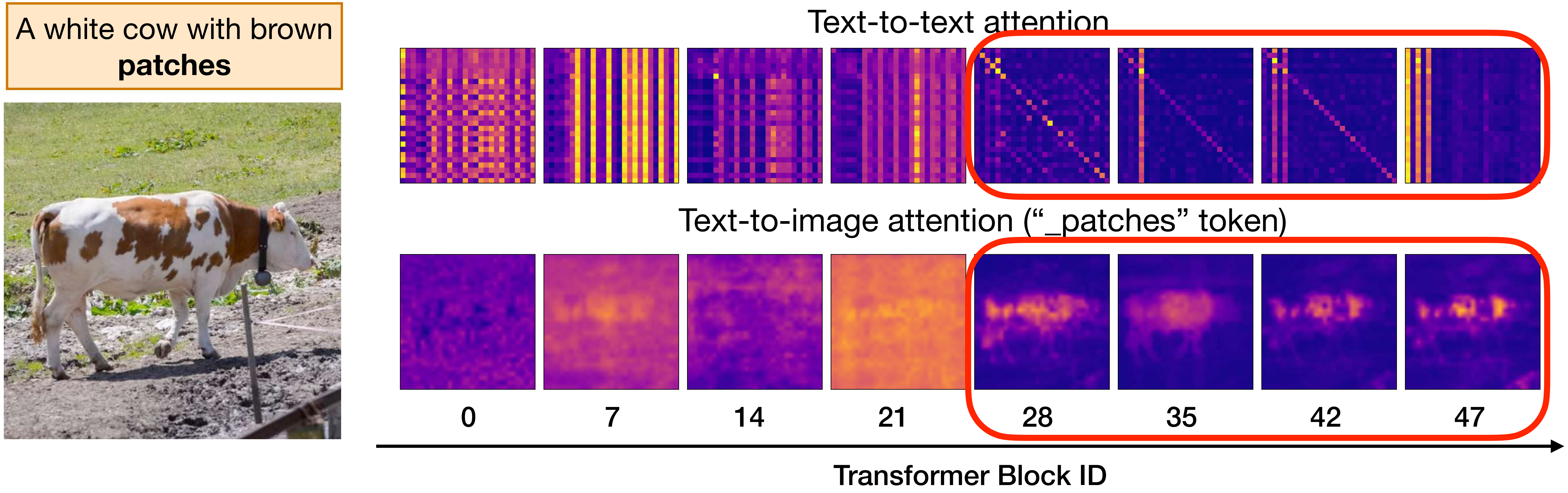


Emergence of Semantic Information in DiT



Early Layers: No usable information, uniform attention maps

Emergence of Semantic Information in DiT

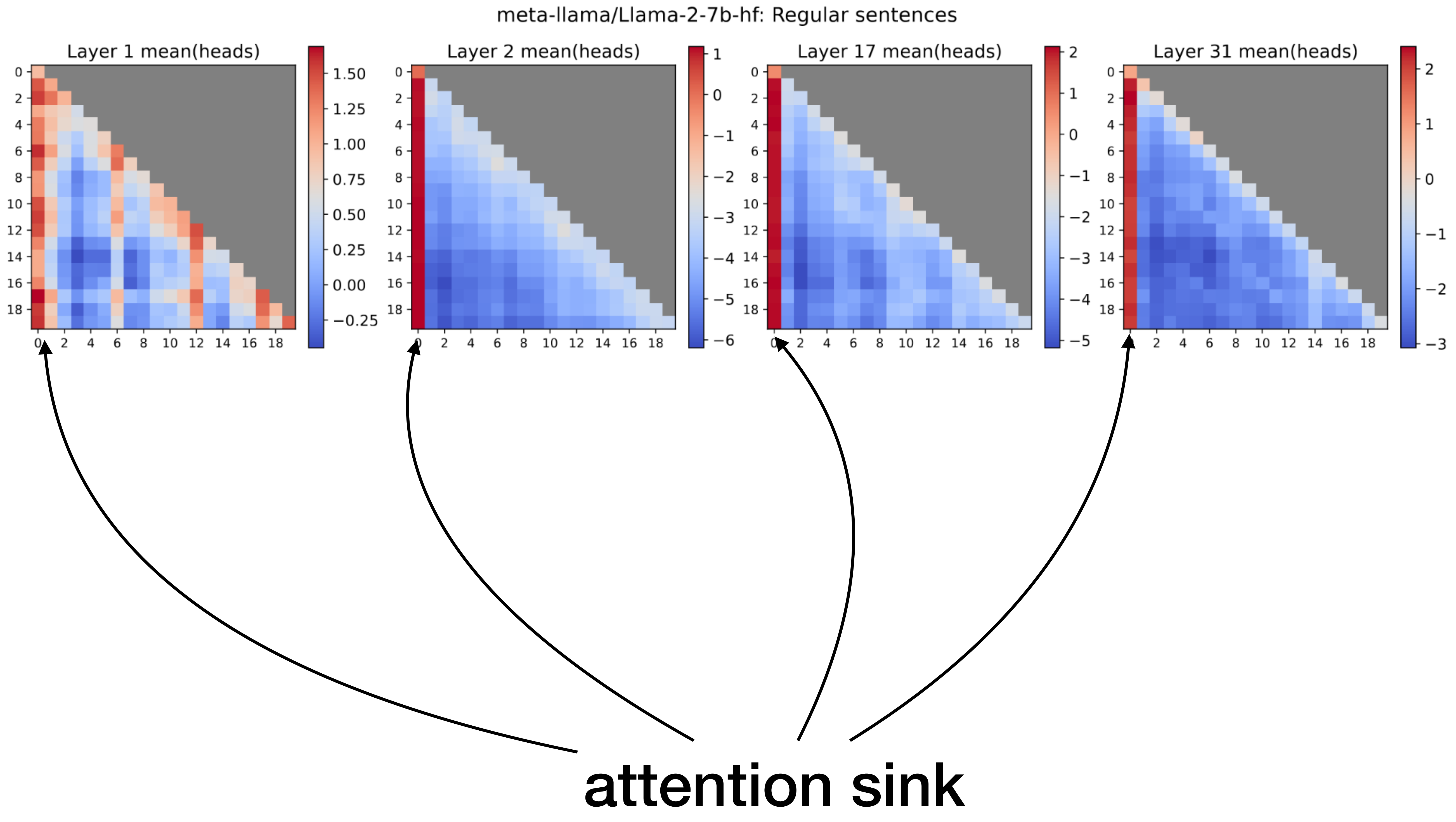


Mid & Late Layers: Sharpened semantic alignment + global attention sinks

Attention Sinks in NLP [1,2,3,4,5] and vision [6,7]

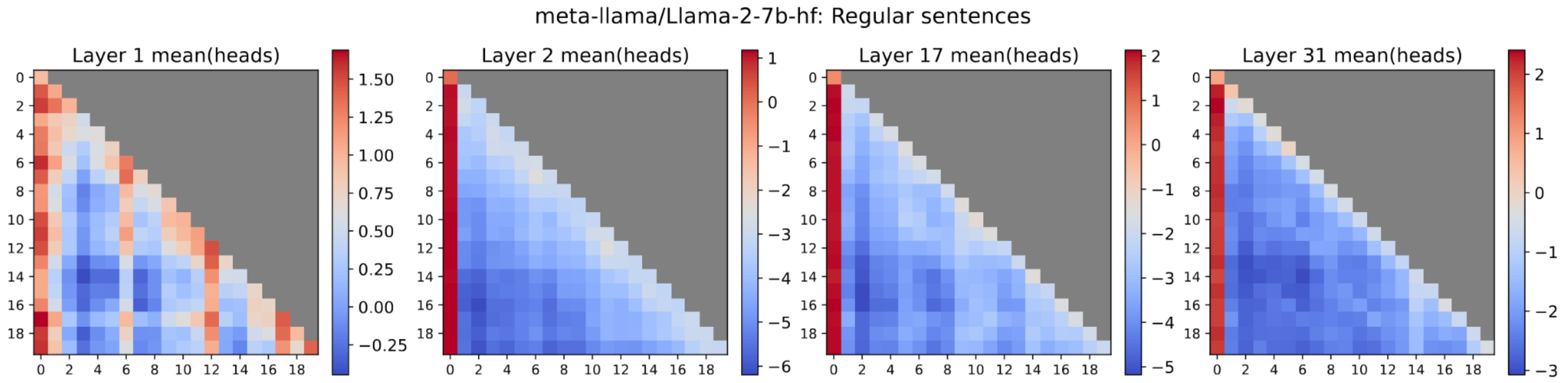
What is attention sink?

- high-norm values
- limited semantic information
- very few tokens



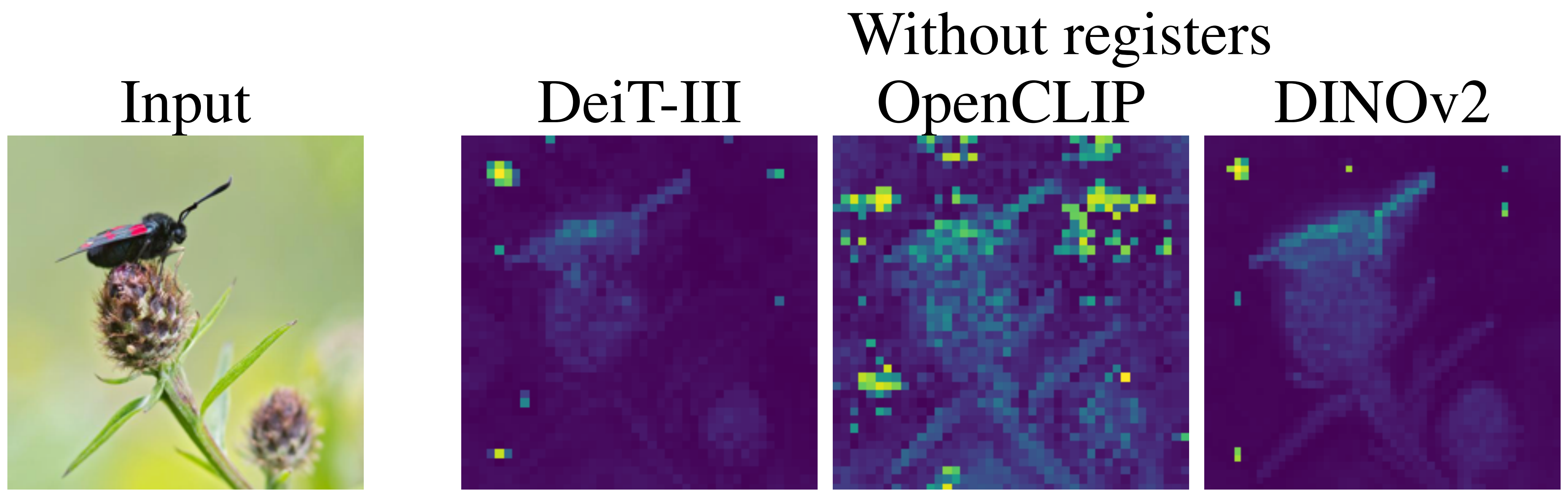
[1] Efficient Streaming Language Models with Attention Sinks, ICLR 2024
[2] Interpreting the Repeated Token Phenomenon in Large Language Models, ICML 2025
[3] Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding, ICML 2025
[4] Massive Activations in Large Language Models, CoLM 2024
[5] Why do LLMs attention to the first token? arxiv 2025
[6] Vision Transformers Need Registers, ICLR 2024
[7] Vision Transformers Don't Need Registers, arxiv 2025

Attention Sinks in NLP [1,2,3,4,5] and vision [6,7]



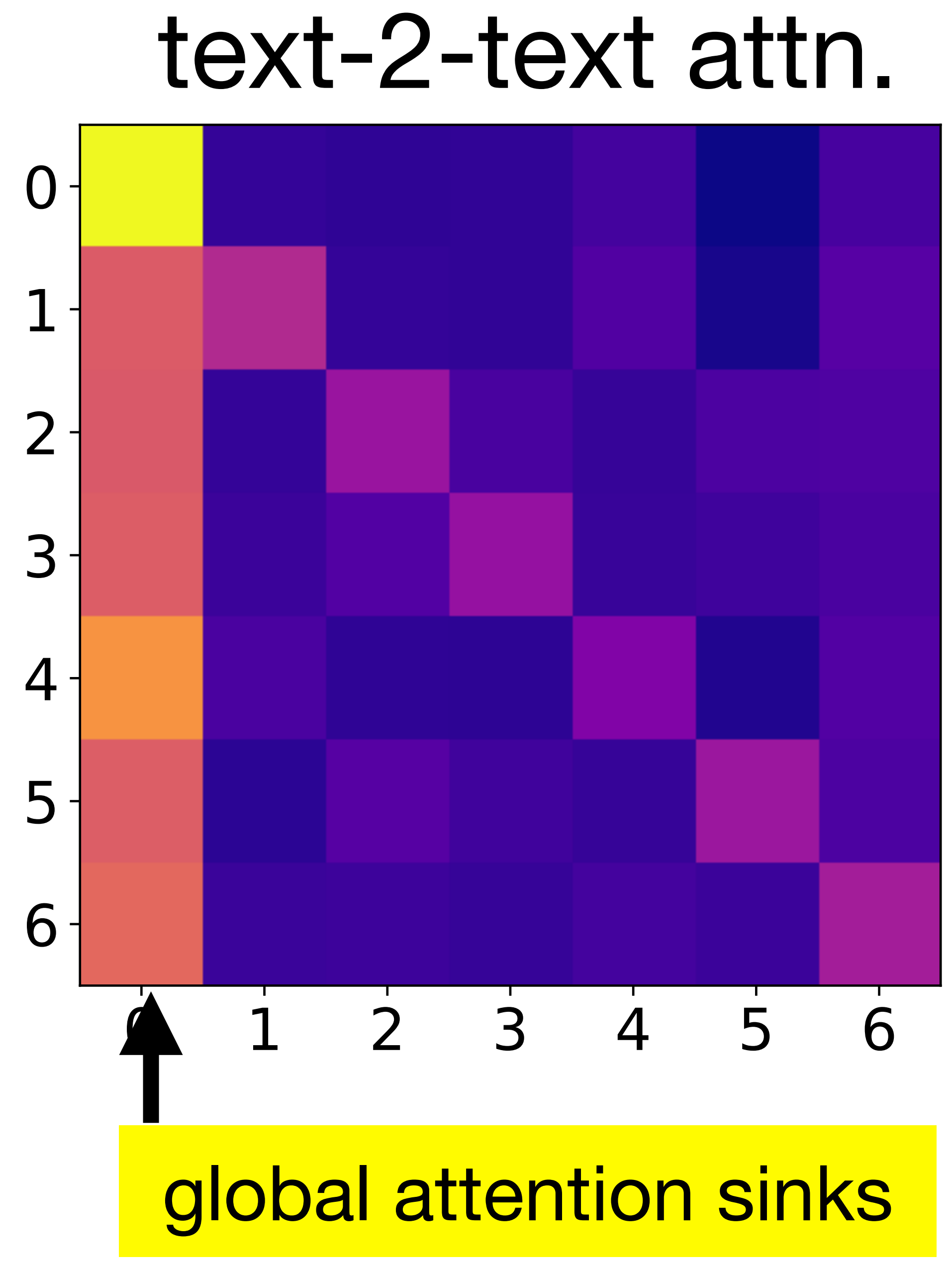
What is attention sink?

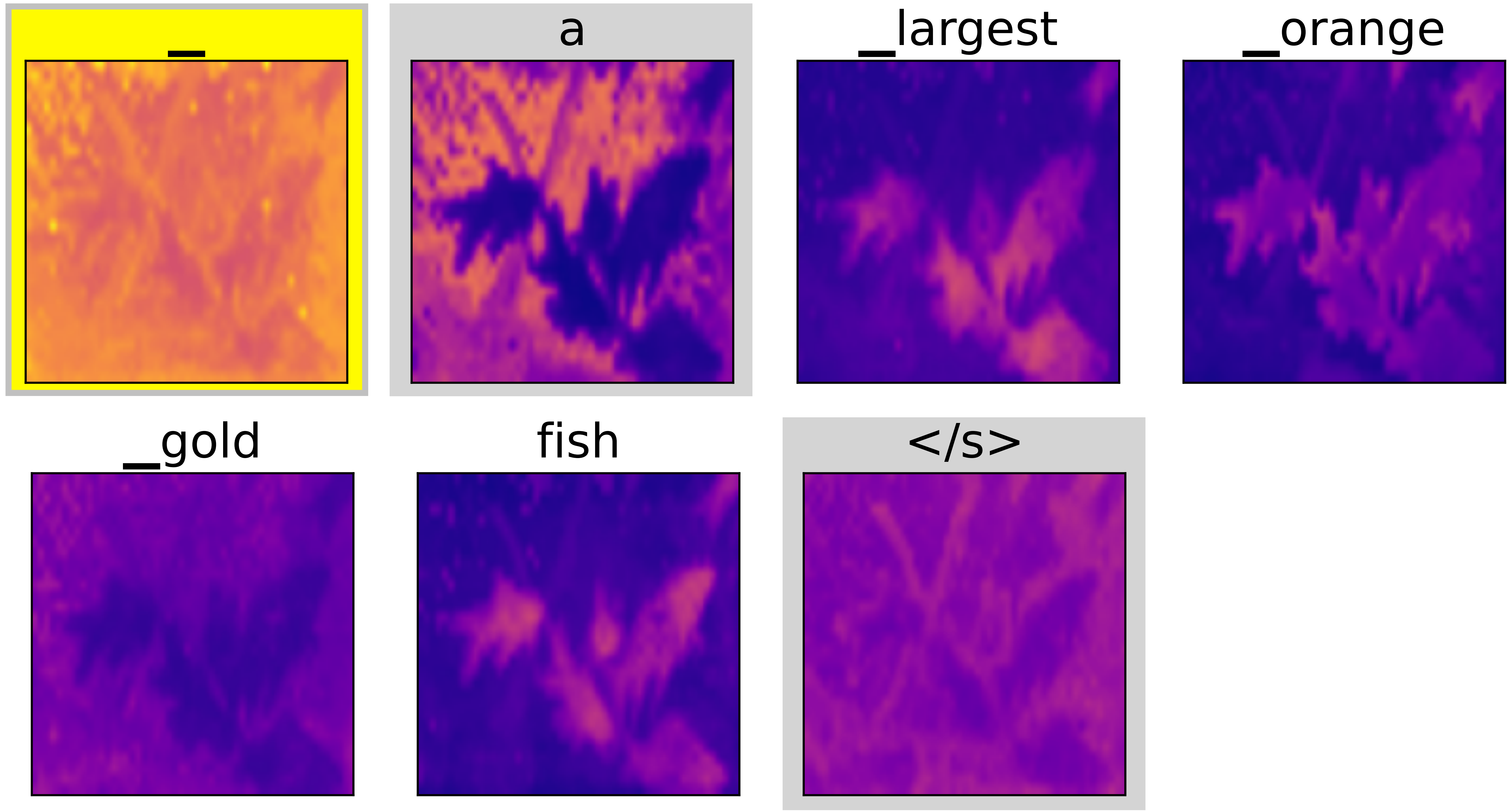
- high-norm values
- limited semantic information
- very few tokens



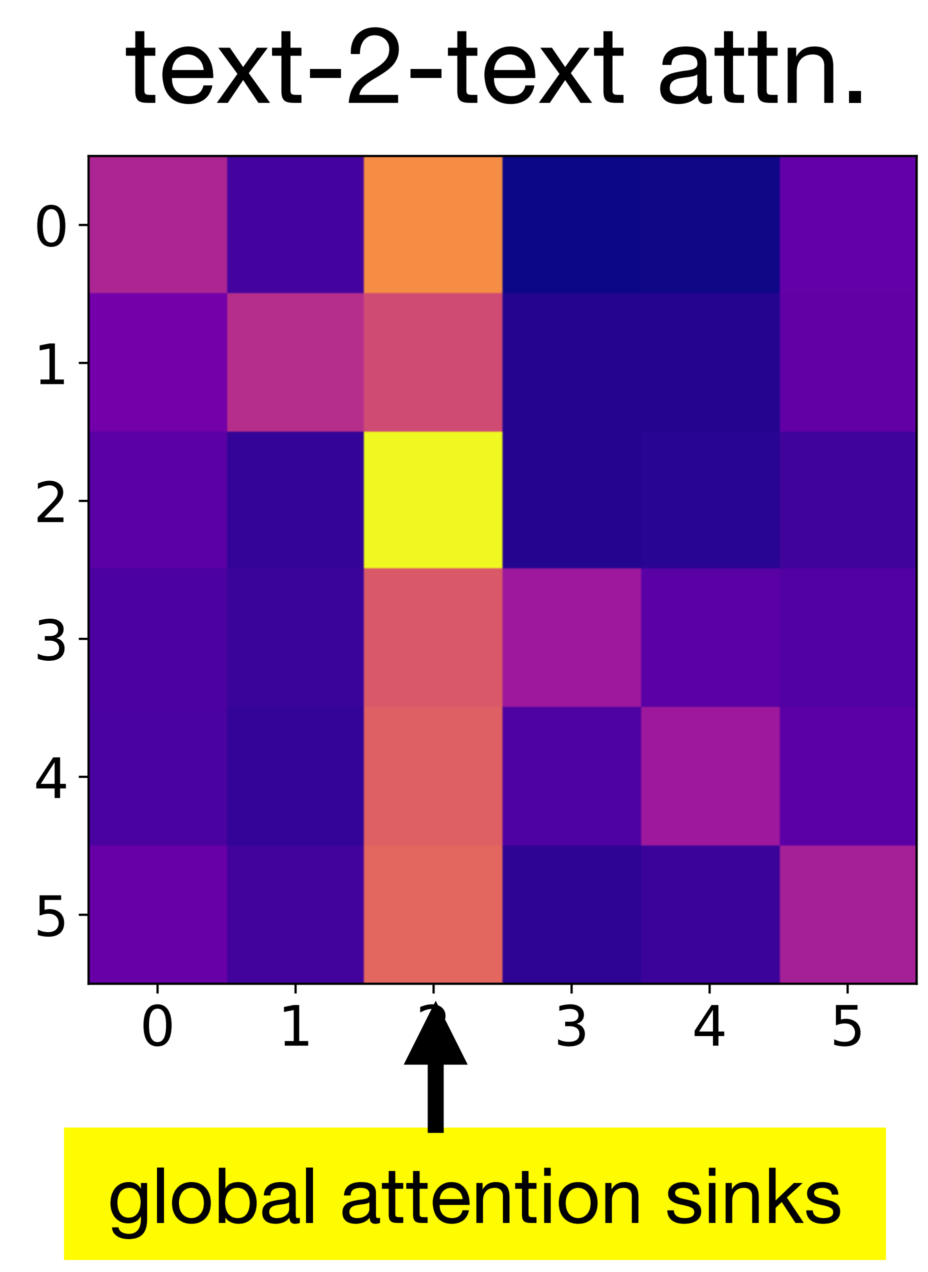
[1] Efficient Streaming Language Models with Attention Sinks, ICLR 2024
[2] Interpreting the Repeated Token Phenomenon in Large Language Models, ICML 2025
[3] Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding, ICML 2025
[4] Massive Activations in Large Language Models, CoLM 2024
[5] Why do LLMs attention to the first token? arxiv 2025
[6] Vision Transformers Need Registers, ICLR 2024
[7] Vision Transformers Don’t Need Registers, arxiv 2025

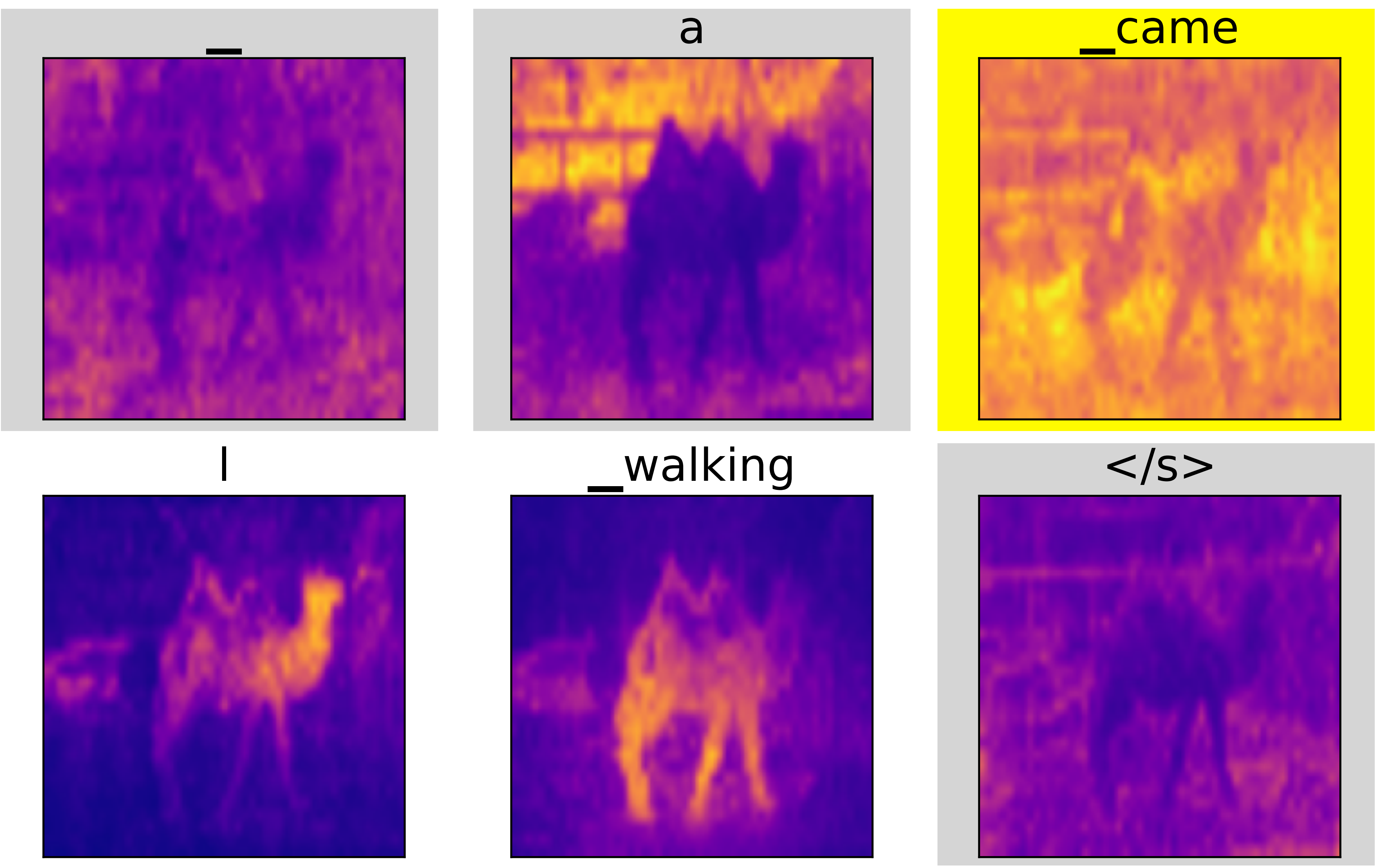
Global Attention Sinks (GAS)





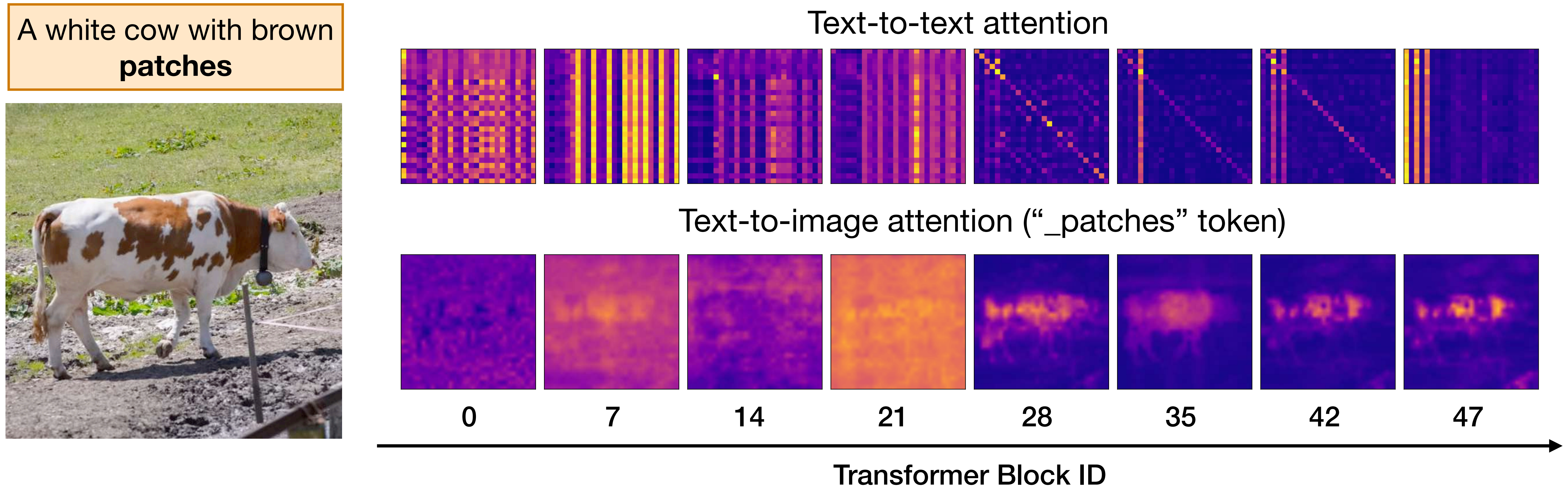
Global Attention Sinks (GAS)





Meaningful token is allocated to GAS

Interpretation of GAS



1. **Uninformative role:** Removing them does not harm the performance (inference)
2. **Indicators of semantic structure:** GAS consistently emerge only after meaningful structure is established in the mid layers
3. **Potentially harmful role:** majority of GAS tokens (77%) correspond to stop words, 10% fall on color tokens and another 10% to other content words

Redistribution Strategy

Append more stop words (attention magnets)!

Redistribution Strategy

Append more stop words (attention magnets)!

stop words: the, is, at, which, on, with, to, a, this, etc

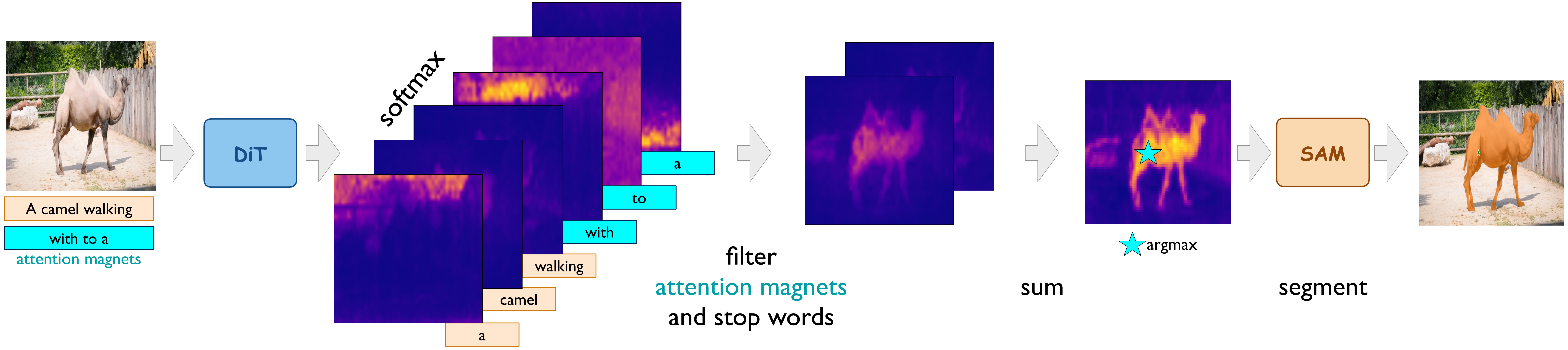
words with little semantic value

Redistribution Strategy

Append more stop words (attention magnets)!

stop words: the, is, at, which, on, with, to, a, this, etc

words with little semantic value



Redistribution Strategy with Attention Magnets

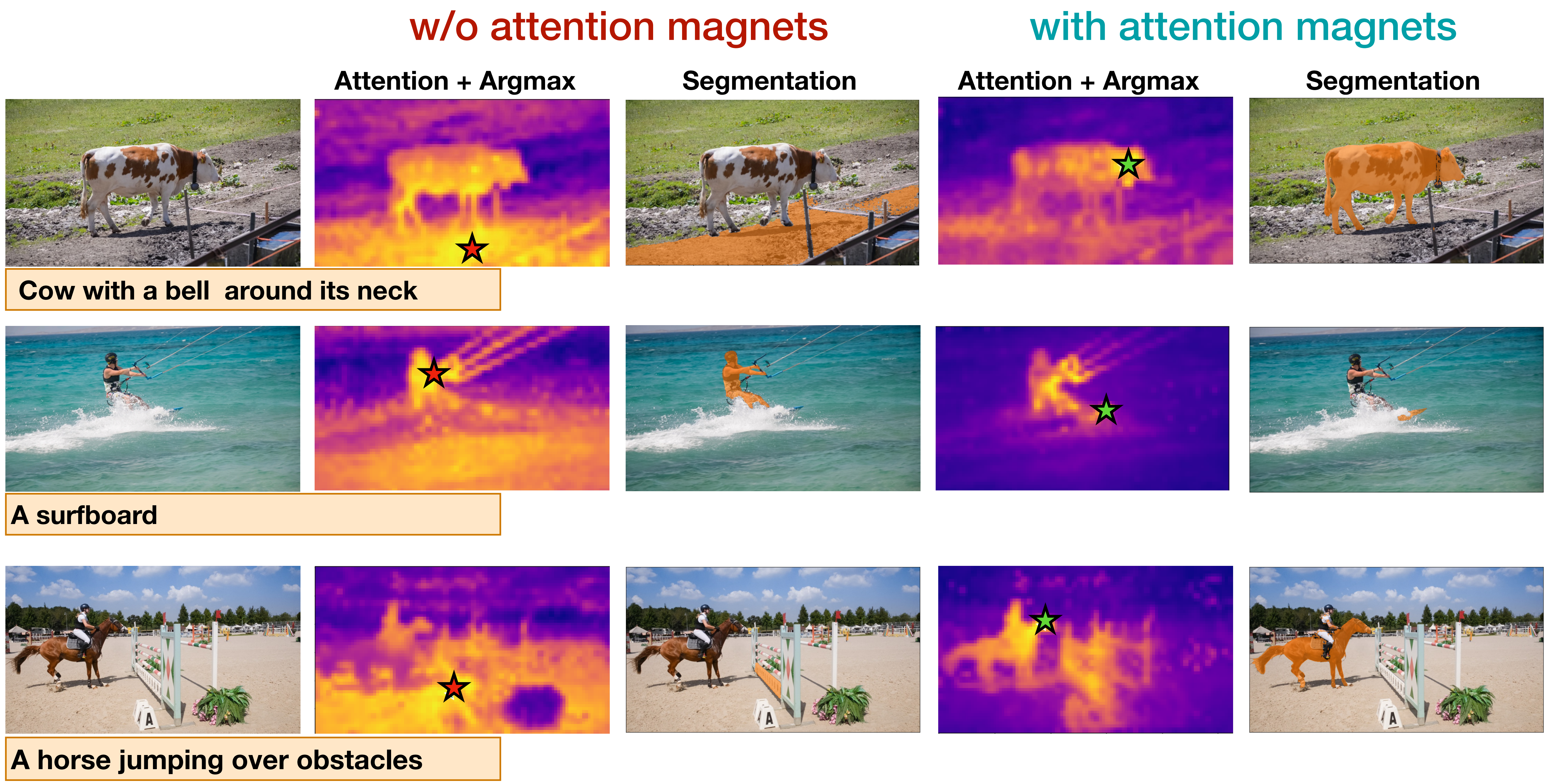
before: 77% of GAS tokens on stop words

after: 89% of GAS tokens on stop words

Redistribution Strategy with Attention Magnets

before: 77% of GAS tokens on stop words

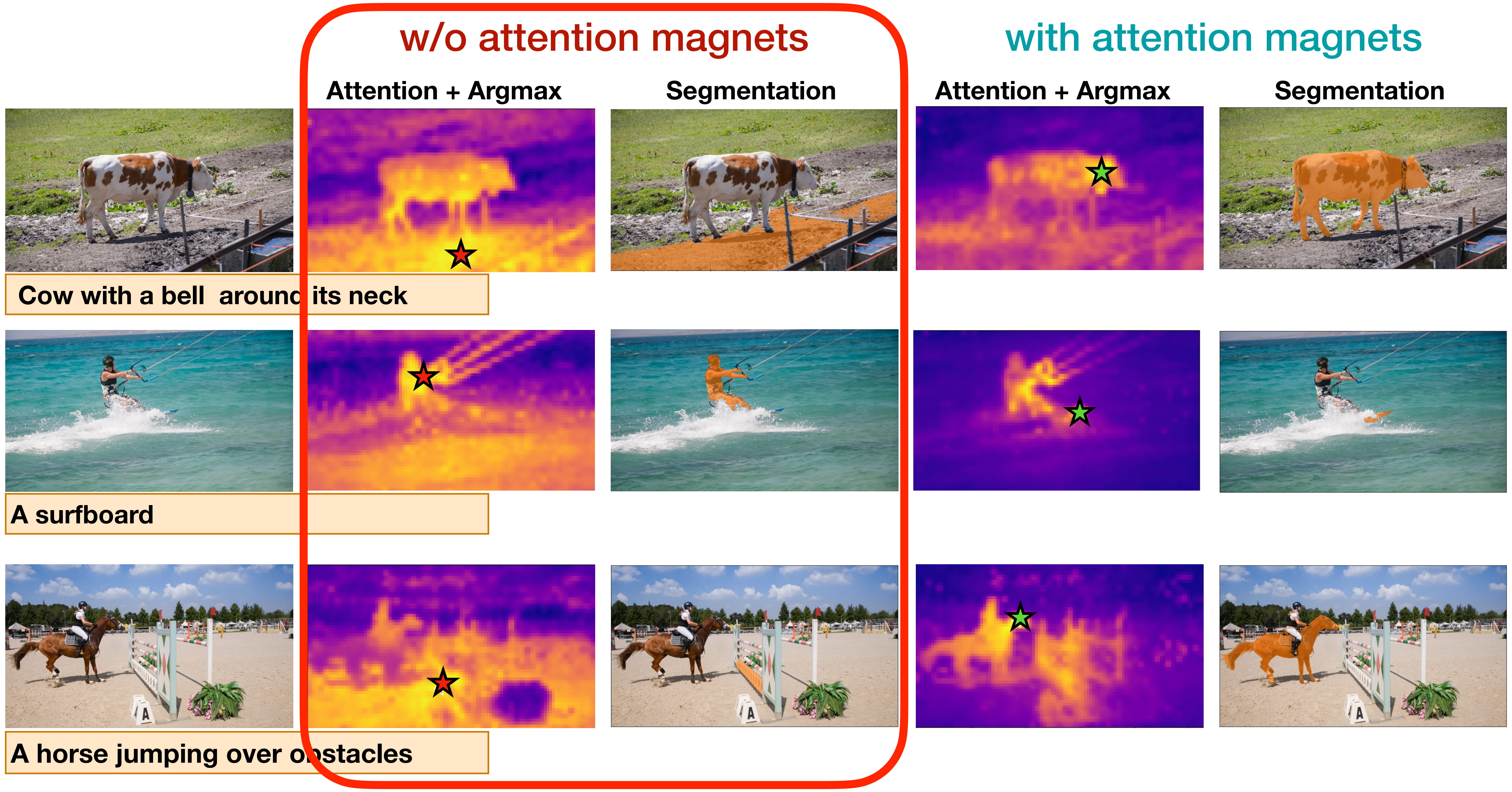
after: 89% of GAS tokens on stop words



Redistribution Strategy with Attention Magnets

before: 77% of GAS tokens on stop words

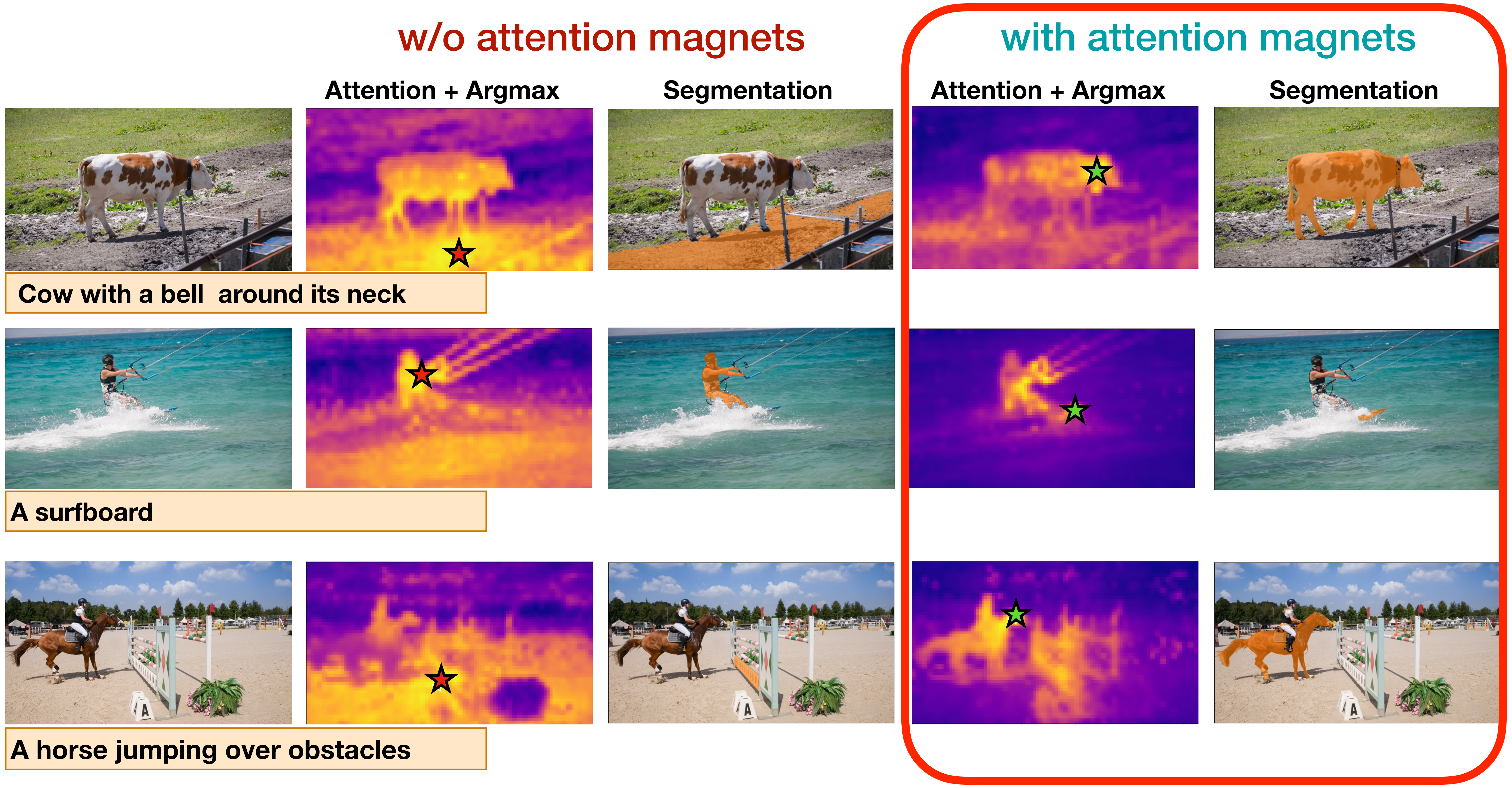
after: 89% of GAS tokens on stop words



Redistribution Strategy with Attention Magnets

before: 77% of GAS tokens on stop words

after: 89% of GAS tokens on stop words



much sharper attention maps

Why Stop Words?

- natural garbage collectors in LLMs → allocation of the surplus of attention
- background attention redistributed to these stop words

Why Stop Words?

- natural garbage collectors in LLMs → allocation of the surplus of attention
- background attention redistributed to these stop words

- is the choice of stop words important?

AM	Ref-DAVIS17 $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	PA
random stop words (5x)	57.5	54.3	60.5	68.5
random vectors (5x)	56.2	53.1	59.4	65.5
none	54.4	50.9	57.6	59.8
scene description	48.9	45.2	52.2	60.6

SOTA

Metric	Method	Vision Backbone	Pre-trained Model	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test
oIoU	*zero-shot methods w/ additional training*										
	Pseudo-RIS (Yu et al., 2024)	ViT-B	SAM, CoCa, CLIP	37.33	43.43	31.90	40.19	46.43	33.63	41.63	43.52
	VLM-VG (Wang et al., 2025)	R101	COCO*, VLM-VG*	45.40	48.00	41.40	37.00	40.70	30.50	42.80	44.10
	zero-shot methods w/o additional training										
	Grad-CAM (Selvaraju et al., 2017a)	R50	SAM, CLIP	23.44	23.91	21.60	26.67	27.20	24.84	23.00	23.91
	MaskCLIP (Zhou et al., 2022)	R50	SAM, CLIP	20.18	20.52	21.30	22.06	22.43	24.61	23.05	23.41
	Global-Local (Yu et al., 2023)	R50	FreeSOLO, CLIP	24.58	23.38	24.35	25.87	24.61	25.61	30.07	29.83
	Global-Local (Yu et al., 2023)	R50	SAM, CLIP	24.55	26.00	21.03	26.62	29.99	22.23	28.92	30.48
	Global-Local (Yu et al., 2023)	ViT-B	SAM, CLIP	21.71	24.48	20.51	23.70	28.12	21.86	26.57	28.21
	Ref-Diff (Ni et al., 2023)	ViT-B	SAM, SD, CLIP	35.16	37.44	34.50	35.56	38.66	31.40	38.62	37.50
	TAS (Suo et al., 2023)	ViT-B	SAM, BLIP2, CLIP	29.53	30.26	28.24	33.21	38.77	28.01	35.84	36.16
	HybridGL (Liu & Li, 2025)	ViT-B	SAM,CLIP	41.81	44.52	38.50	35.74	41.43	30.90	42.47	42.97
	REFAM (ours)	DiT	SAM, FLUX	**46.91**	**52.30**	**43.88**	**38.57**	**42.66**	**34.90**	**45.53**	**44.45**

Referral Image Object Segmentation

Method	Ref-DAVIS17 $\mathcal{J}\&\mathcal{F}$	Ref-DAVIS17 $\mathcal{J}$	Ref-DAVIS17 $\mathcal{F}$	Ref-YouTube-VOS $\mathcal{J}\&\mathcal{F}$	Ref-YouTube-VOS $\mathcal{J}$	Ref-YouTube-VOS $\mathcal{F}$	MeViS $\mathcal{J}\&\mathcal{F}$	MeViS $\mathcal{J}$	MeViS $\mathcal{F}$
Training-Free with Grounded-SAM									
Grounded-SAM (Ren et al., 2024)†	65.2	62.3	68.0	62.3	61.0	63.6	-	-	-
Grounded-SAM2 (Ren et al., 2024)†	66.2	62.6	69.7	64.8	62.5	67.0	38.9	35.7	42.1
AL-Ref-SAM2 (Huang et al., 2025)	74.2	70.4	78.0	67.9	65.9	69.9	42.8	39.5	46.2
Training-Free									
G-L + SAM2 (Yu et al., 2023)†	40.6	37.6	43.6	27.0	24.3	29.7	23.7	20.4	30.0
G-L (SAM) + SAM2 (Yu et al., 2023)†	46.9	44.0	49.7	33.6	29.9	37.3	26.6	22.7	30.5
REFAM + SAM2 (ours)	**57.6**	**54.5**	**60.6**	**42.7**	**37.6**	**47.8**	**30.6**	**24.7**	**36.6**

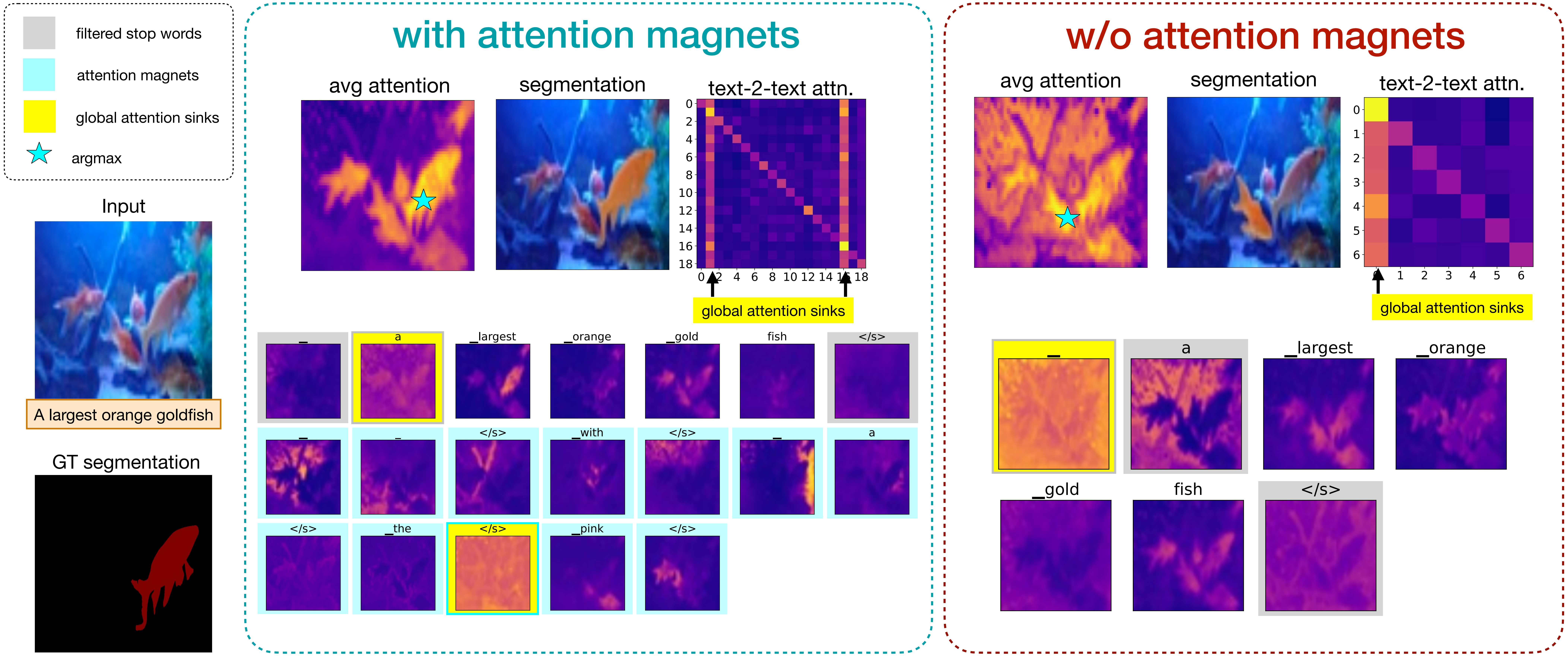
Referral Video Object Segmentation

Does our redistribution strategy help?

AM	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
✔	46.91	52.30	43.88	38.57	42.66	34.90	45.53	44.45
-	33.89	44.66	34.14	35.12	37.69	33.75	42.93	42.44

With and Without Attention Magnets (AM)

Qualitative Example



Conclusion

- RefAM framework for zero-shot referral segmentation based on DiT
- Step forward in understanding semantics in diffusion models through the lens of LLMs
- Attention redistribution strategy with attention magnets
- SOTA results on zero-shot image an video referral segmentation

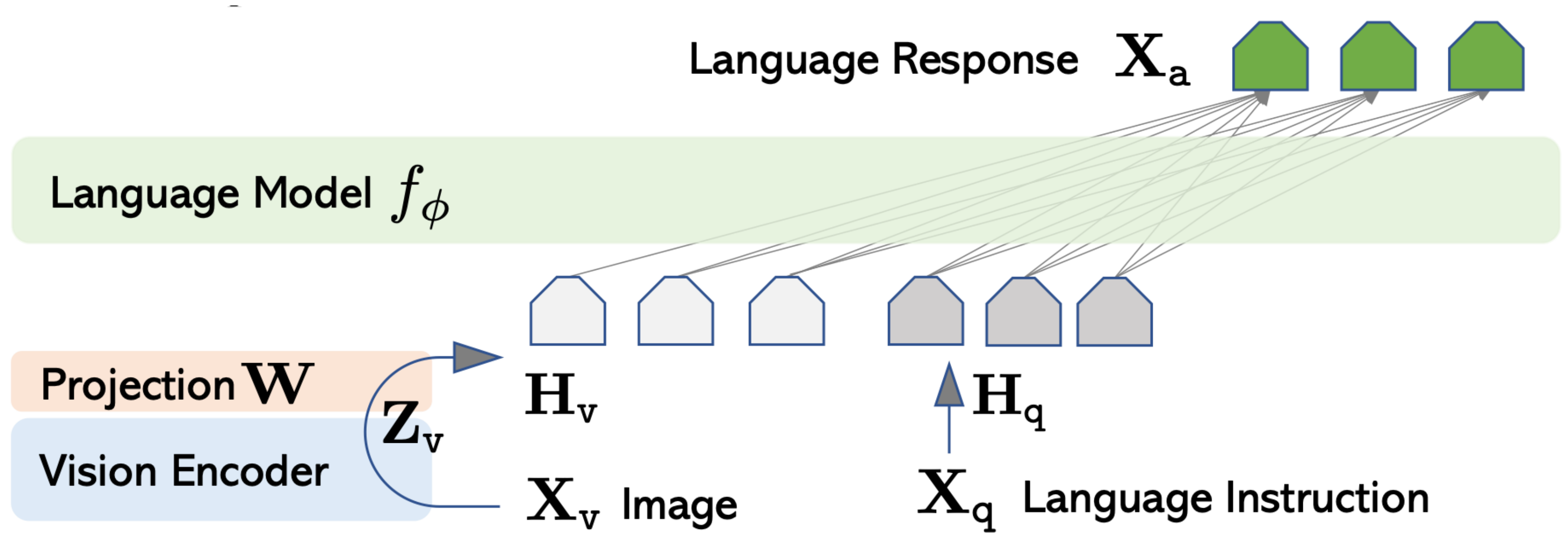
Language-Unlocked ViT (LUViT): Empowering Self-Supervised ViT with LLMs

Selim Kuzucu[1], Ferjad Naeem[2], Anna Kukleva[1], Federico Tombari[2,3], Bernt Schiele[1]

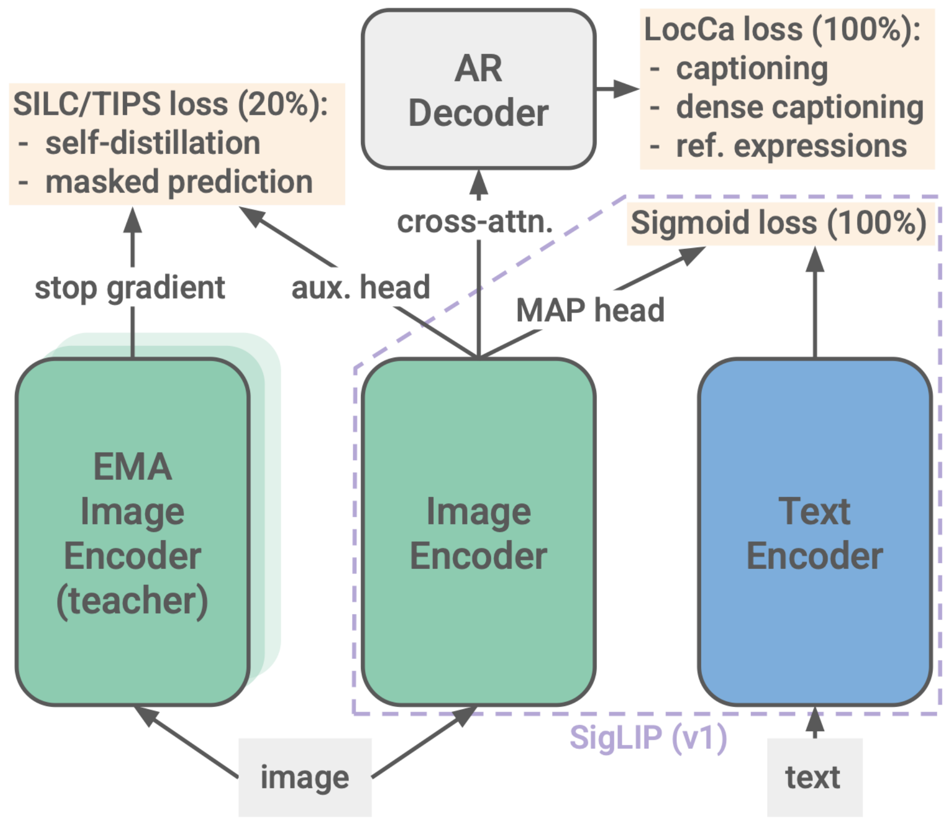
[1]Max Planck Institute for Informatics, [2]Google, [3]TU Munich

Leveraging pre-trained LLM representations for pure vision tasks

Pretrained LLMs in vision



LLaVa [1]

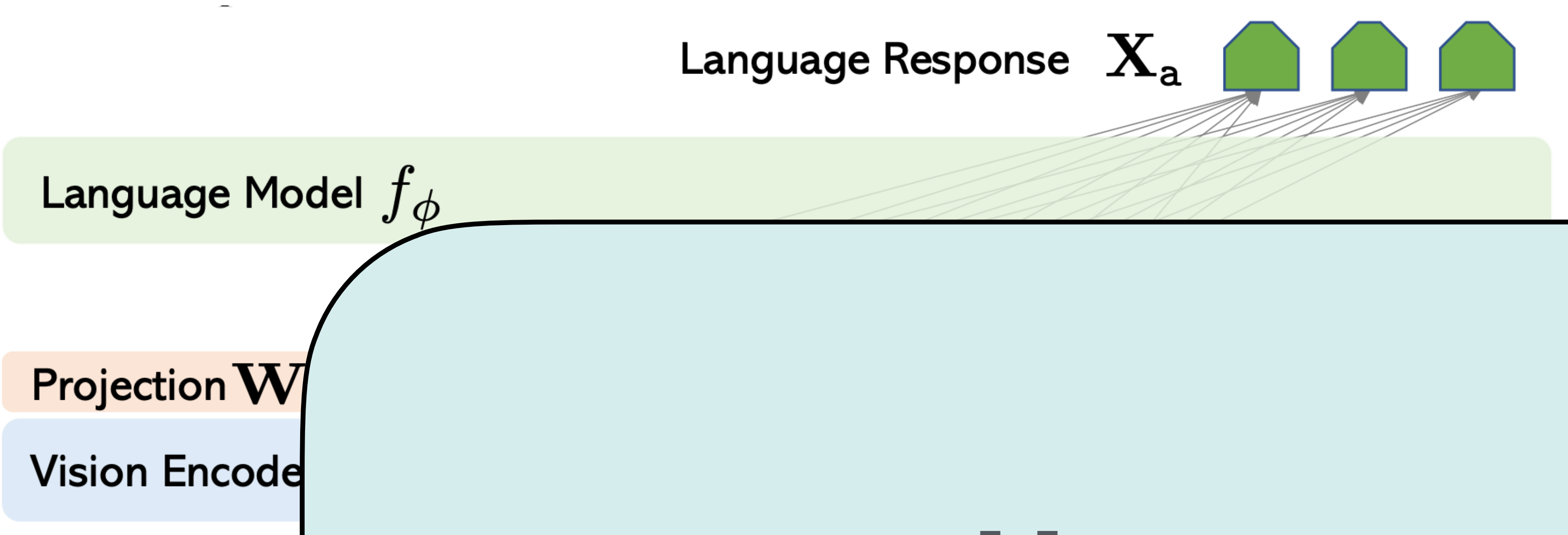


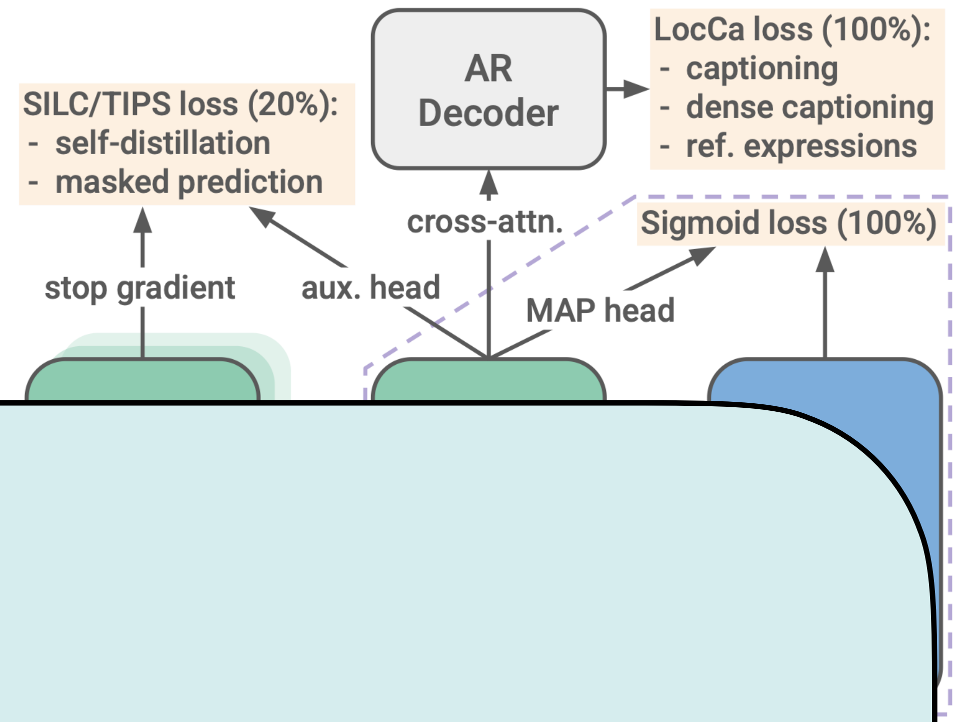
SigLIP 2 [2]

LLMs can process visual information..
IF they are trained jointly with visual encoders on vast data!

[1] Visual Instruction Tuning, NeurIPS 2023
[2] SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, arxiv

Pretrained LLMs in vision





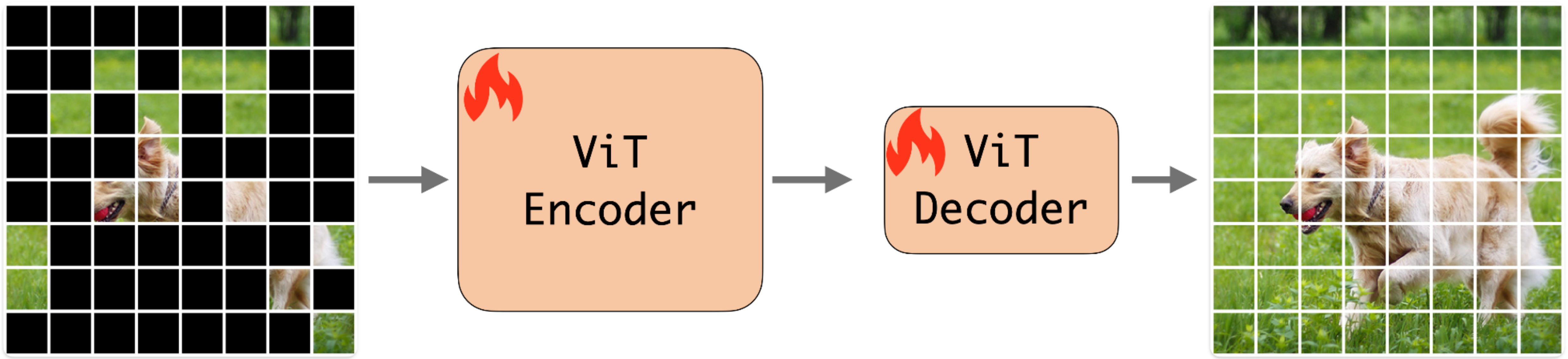
How can we improve ViT with an off-the-shelf LLM?

IF they are trained jointly with visual encoders on vast data!

[1] Visual Instruction Tuning, NeurIPS 2023
[2] SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, arxiv

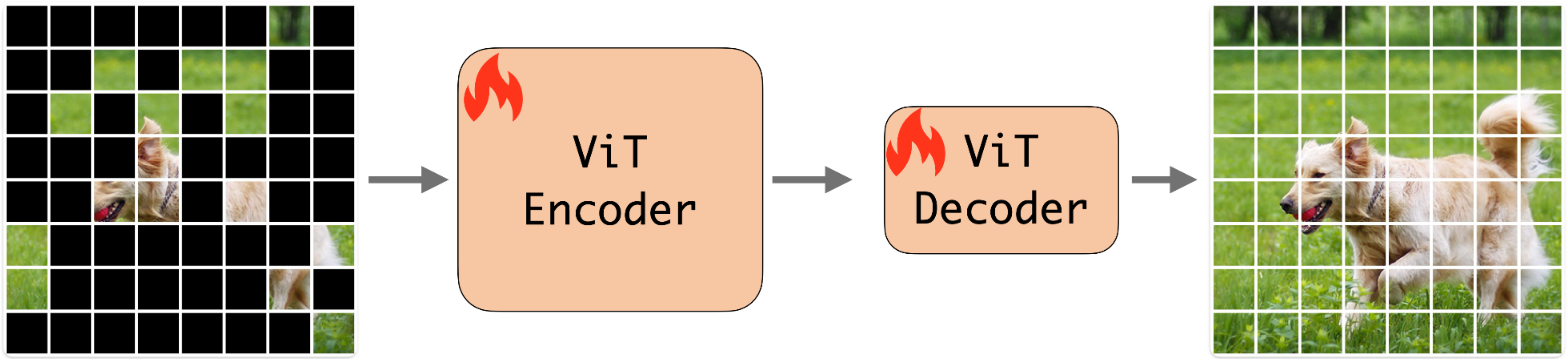
Language-unlocked ViT (LUViT)

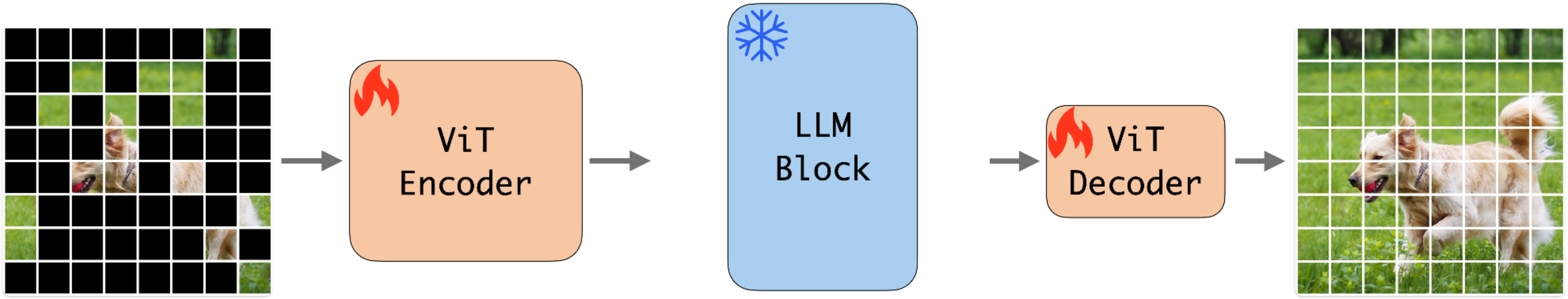
Masked Autoencoder (MAE)



Language-unlocked ViT (LUViT)

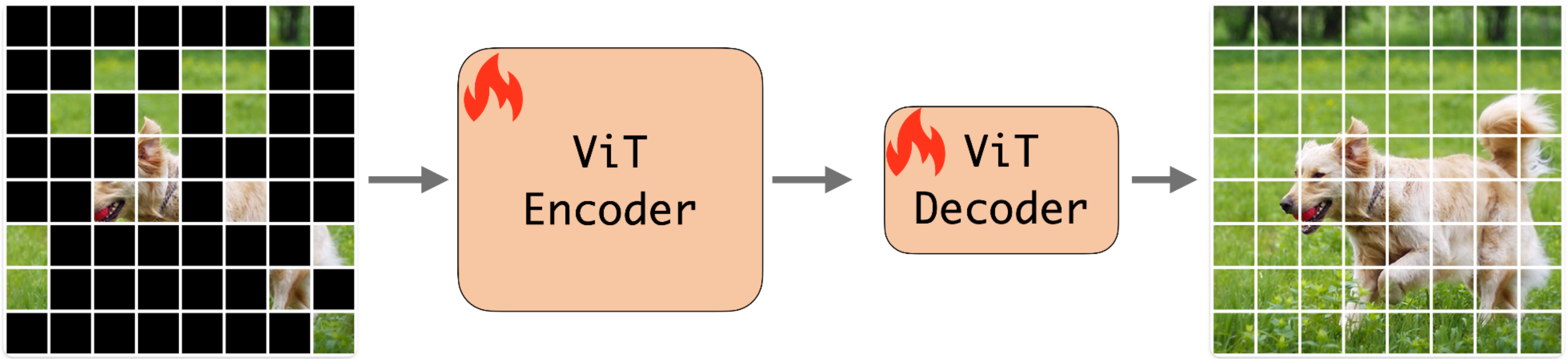
Masked Autoencoder (MAE)

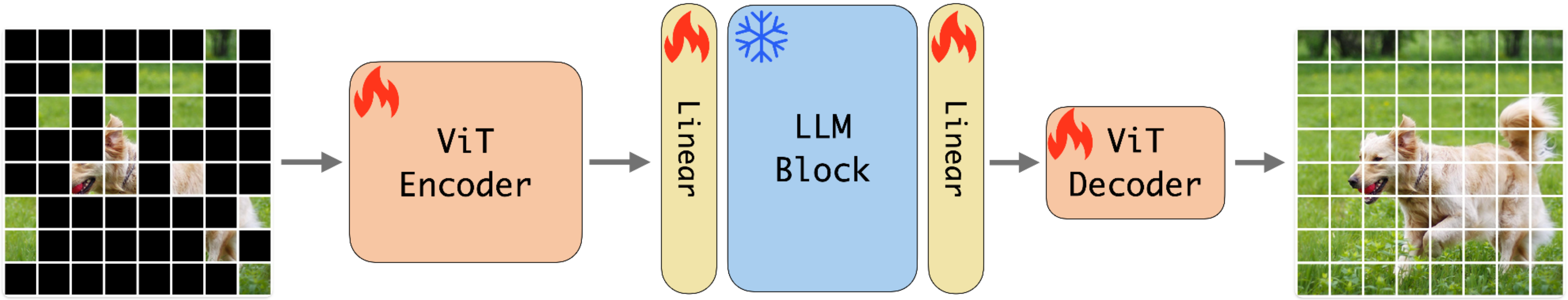




Language-unlocked ViT (LUViT)

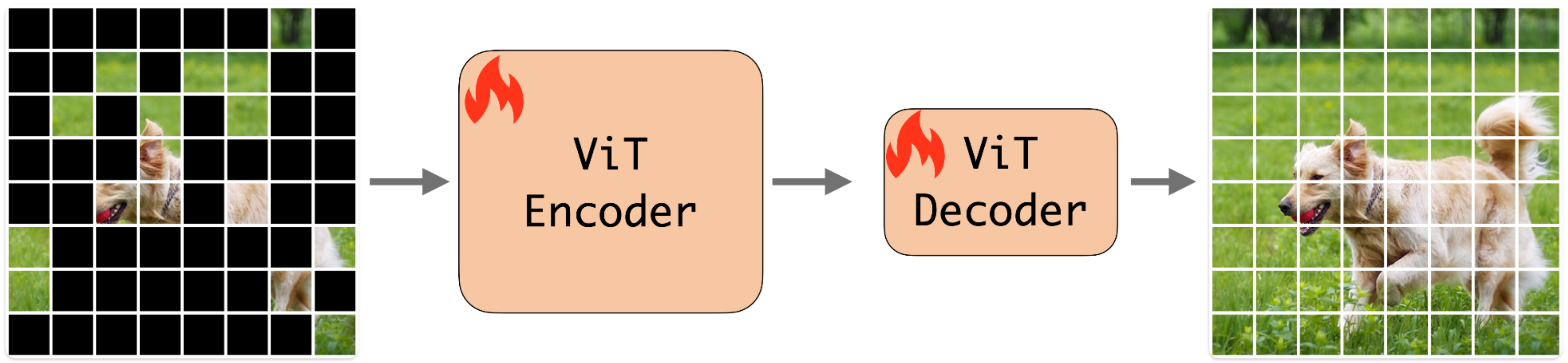
Masked Autoencoder (MAE)



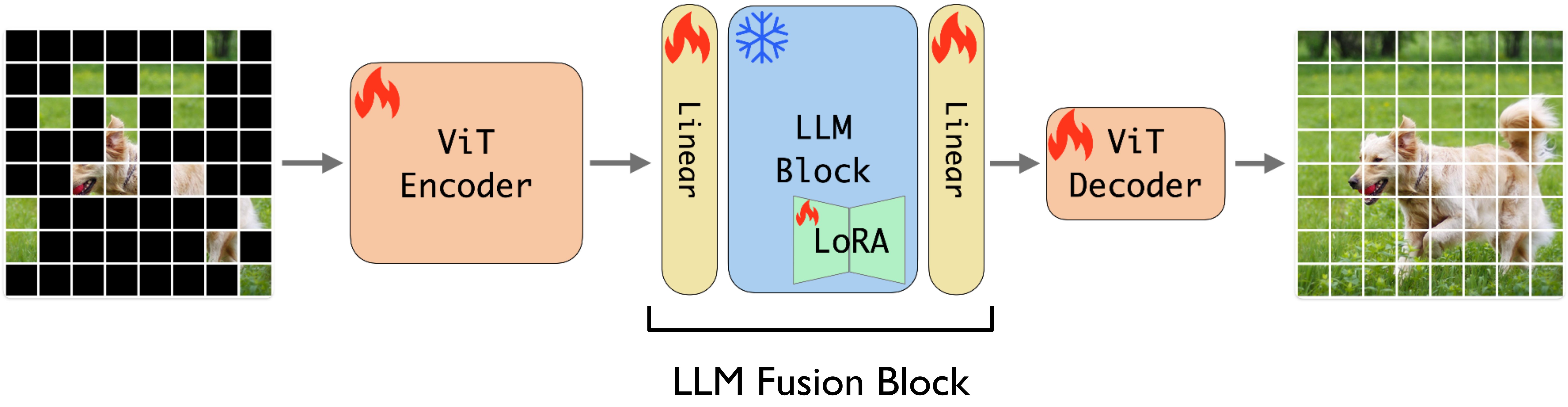


Language-unlocked ViT (LUViT)

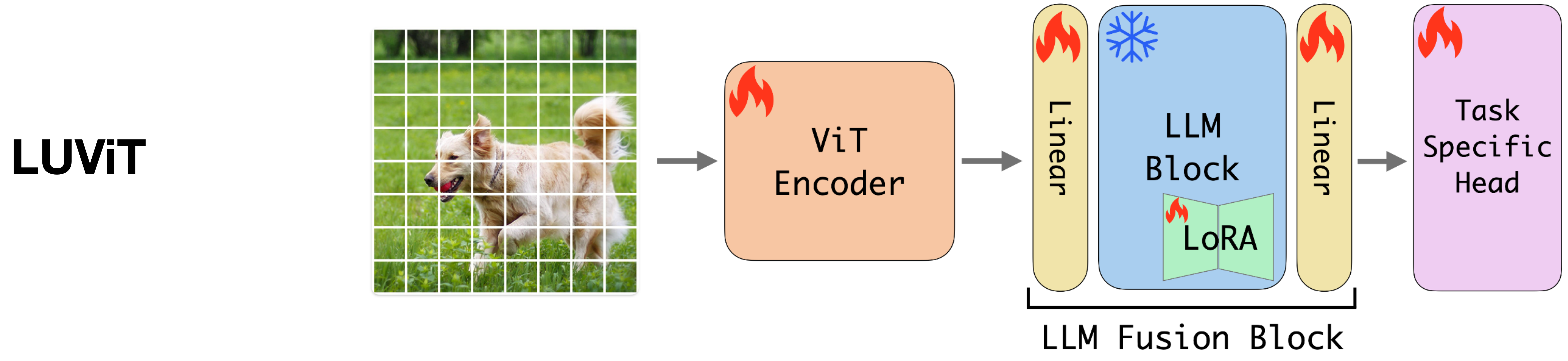
Masked Autoencoder (MAE)



LUViT



Language-unlocked ViT (LUViT)



- Effective adaptation of the LLM for pretraining/finetuning
- Single MAE objective for training both ViT and LoRA parameters

Discriminative Visual Tasks

Training	Model	IN-1K	IN-A	IN-SK	IN-V2	IN-R	IN-C
MAE Pretrained	ViT/B	$83.11_{\pm 0.09}$	$33.64_{\pm 0.11}$	$35.69_{\pm 0.30}$	$72.73_{\pm 0.21}$	$49.88_{\pm 0.32}$	$62.86_{\pm 0.01}$
	LUViT *(Ours)*	$\mathbf{83.63}_{\pm \mathbf{0.04}}$	$\mathbf{36.39}_{\pm \mathbf{0.28}}$	$\mathbf{36.36}_{\pm \mathbf{0.61}}$	$\mathbf{73.15}_{\pm \mathbf{0.02}}$	$\mathbf{50.17}_{\pm \mathbf{0.16}}$	$\mathbf{63.44}_{\pm \mathbf{0.05}}$
		+0.52	+2.75	+0.67	+0.42	+0.29	+0.58

ImageNet Classification

Model	Bounding Box			Mask		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
MAE ViT/B	50.6	71.0	55.5	44.9	68.2	48.7
LUViT *(Ours)*	**51.1**	**71.5**	**55.9**	**45.1**	**68.8**	**48.8**
	+0.5	+0.5	+0.4	+0.2	+0.6	+0.1

COCO object detection

It is not just the extra weights!

	Model	Trainable Params.	IN-1K
(a)	ViT/B	86.8M	$83.11_{\pm 0.09}$
(c)	ViT/B+LM1	92.9M	$83.13_{\pm 0.02}$
(f)	LUViT *(Ours)*	93.1M	$\mathbf{83.63_{\pm 0.04}}$

LoRA adaptation is crucial

It is not just the extra weights!

	Model	Trainable Params.	IN-1K
(a)	ViT/B	86.8M	$83.11_{\pm 0.09}$
(e)	ViT/B+Random LM1+LoRA	93.1M	$83.25_{\pm 0.09}$
(f)	LUViT *(Ours)*	93.1M	$\mathbf{83.63}_{\pm \mathbf{0.04}}$

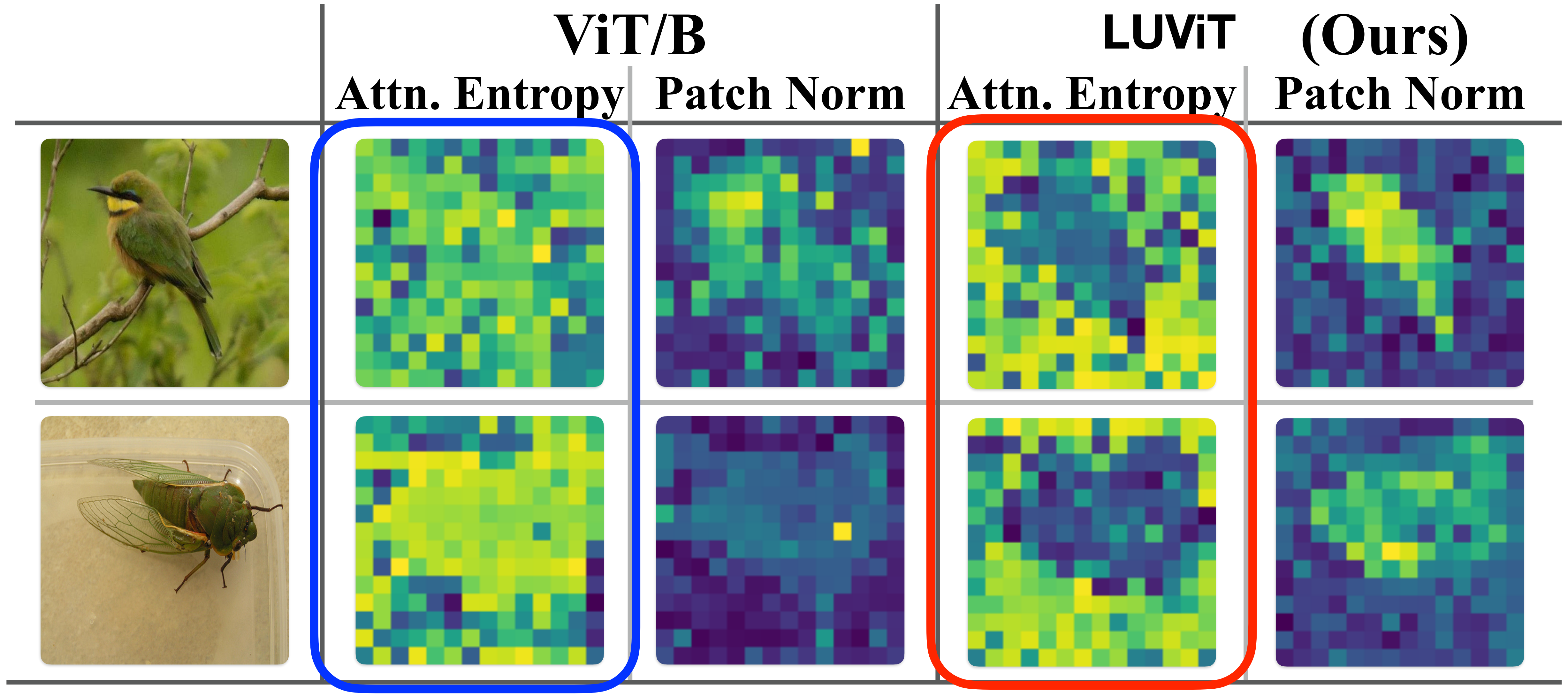
LLM knowledge matter (vs. random parameters with the same # params)

Different LLMs? Different blocks?

		LLM Type	Block	Trainable Params.	IN-1K
MAE ViT/B		N/A	N/A	86.8M	83.2
LUViT	**(a)**	LLaMA 1	1	93.1M	83.2
	(b)	LLaMA 1	16	93.1M	83.4
	(c)	LLaMA 1	31	93.1M	83.5
	(d)	LLaMA 1 (*default*)	32	93.1M	**83.6**
	(e)	Gemma 2	42	93.1M	83.5
	(f)	LLaMA 3.1	32	93.1M	**83.6**
	(g)	LLaMA 3.1-Instruction	32	93.1M	**83.6**

Why does it work?

Background robustness!



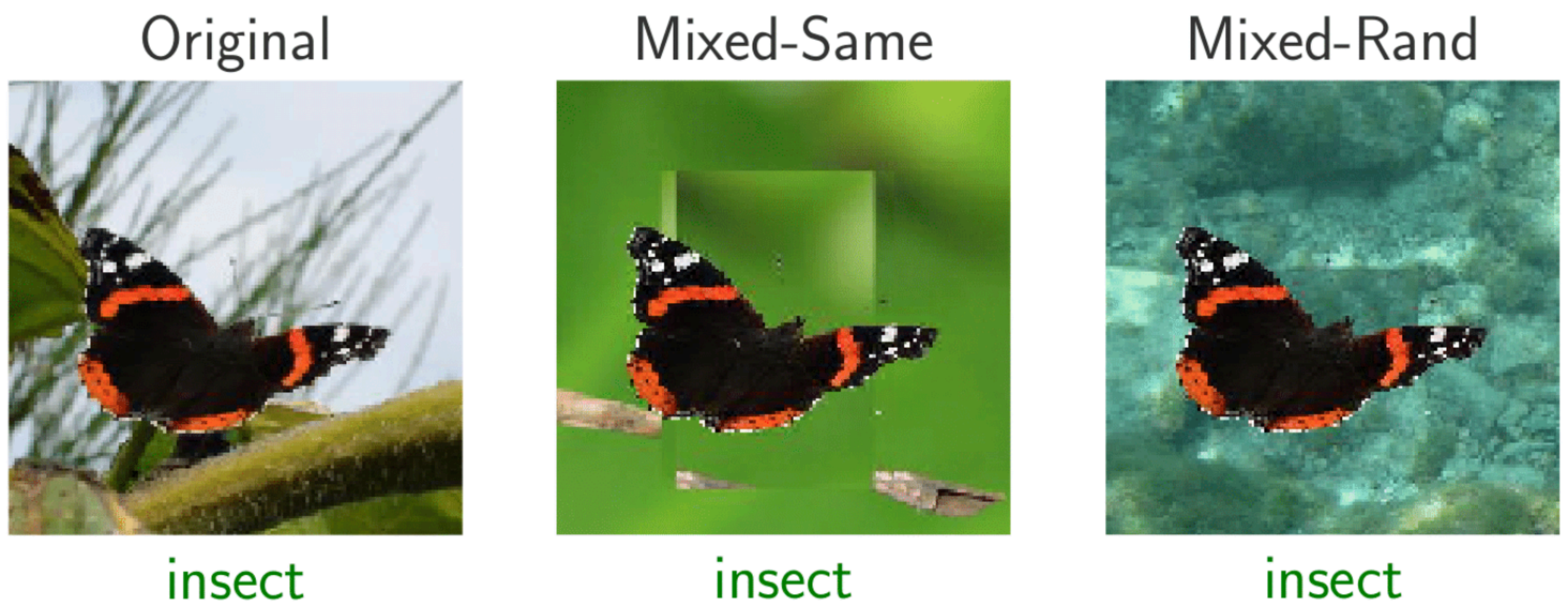
for standard ViT all patches have same attention certainty
whereas LUViT is more certain about foreground (low entropy in dark regions)

Background Robustness

Background Overreliance Benchmark

Image Classification on Imagenet-9

Model	Original	Same	Random	*Orig.-Same*↓	*Orig.-Rand.*↓	*Same-Rand.*↓
MAE ViT/B	96.5	87.8	83.2	8.7	13.3	4.6
LUViT (*Ours*)	**96.6**	**89.2**	**85.3**	**7.4**	**11.3**	**3.9**
	+0.1	+1.4	+2.1	−1.3	−2.0	−0.7



Conclusion

- Pretrained LLMs can be helpful even for purely self-supervised visual representations
- SSL with MAE and LoRA is the recipe to leverage LLMs
- LLM block amplifies informative foreground and attenuates reliance on background

ECCV 2024

HowToCaption: Prompting LLMs to Transform Video Annotations at Scale

Nina Shvetsova[*1,2,3], Anna Kukleva[*1], Xudong Hong[1,4], Christian Rupprecht[5], Bernt Schiele[1], Hilde Kuehne[2,3,6]

[1]Max Planck Institute for Informatics, [2]Goethe University Frankfurt, [3]Bonn University, [4]Saarland University, [5]University of Oxford, [6]MIT-IMB Watson AI Lab

Leveraging pre-trained LLM

for large scale video pretraining

Learning from Web Data (Pretraining)

Narrated Videos

00:15 00:24 00:31 00:35 00:44

ASR subtitles:

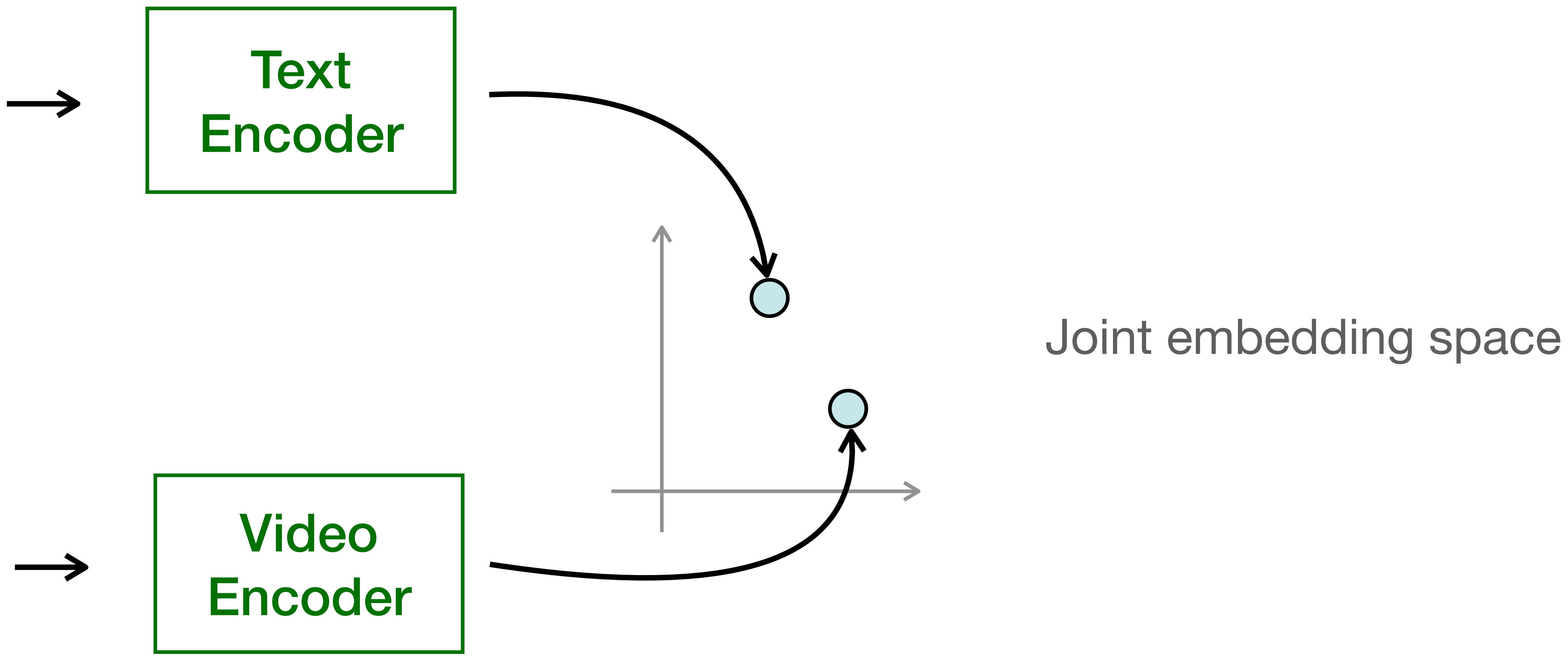
00:15	00:24	00:31	00:35	00:44
so in order to get started we have to have our patient here skeeter my dog and we’re going to get some toothpaste and it's going to be something that she really likes	so this is a chicken flavored toothpaste which she thinks is pretty delightful	okay and then we're just going to get any old toothbrush	they make dog toothbrushes but you can just get a soft children's toothbrush or adult toothbrush for a large dog	so we're going to focus here on the outside edges of the front teeth and the canine teeth

- ✓ Dense textual annotations through ASR narrations
- ✓ Can be collected on a large scale with no human supervision

- – ASR narrations includes noise: incomplete sentences, filler words and phrases, such as “I’m going to”, etc.
- – Alignment of spoken text to the video is very noisy (might be temporal unaligned to video, or completely unrelated)

Narrated Videos for Large-scale Pretraining

6s: In order to get started we have to have our patient skitter my dog here ….
10s: …



Use LLM to transform ASRs into proper aligned captions

HowToCaption Method

ASR + timestamps:

6s: In order to get started we have to have our patient skitter my dog here ….
10s: …

→

LLM

Carefully designed prompt

→

Generated captions:

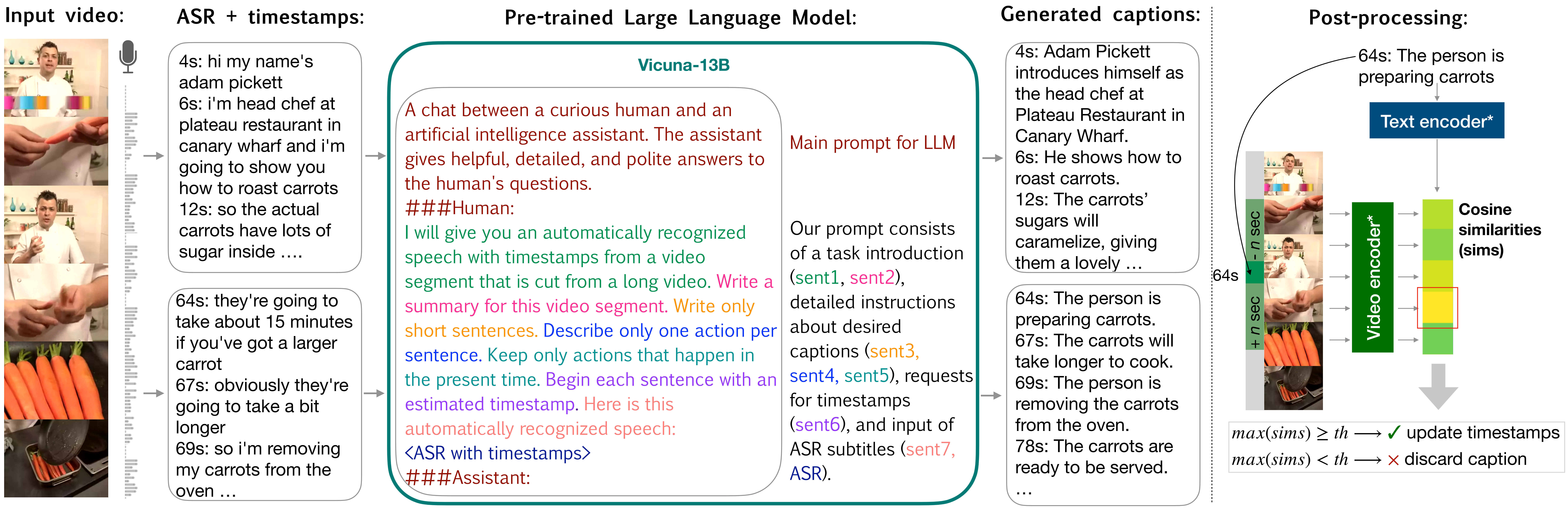
***8s**: Speaker prepares demonstration with the dog and toothpaste*

+

Post-processing

→

HowToCaption — Method



Re-align new captions

HowToCaption — The Dataset

ASR: move them around to help direct the path

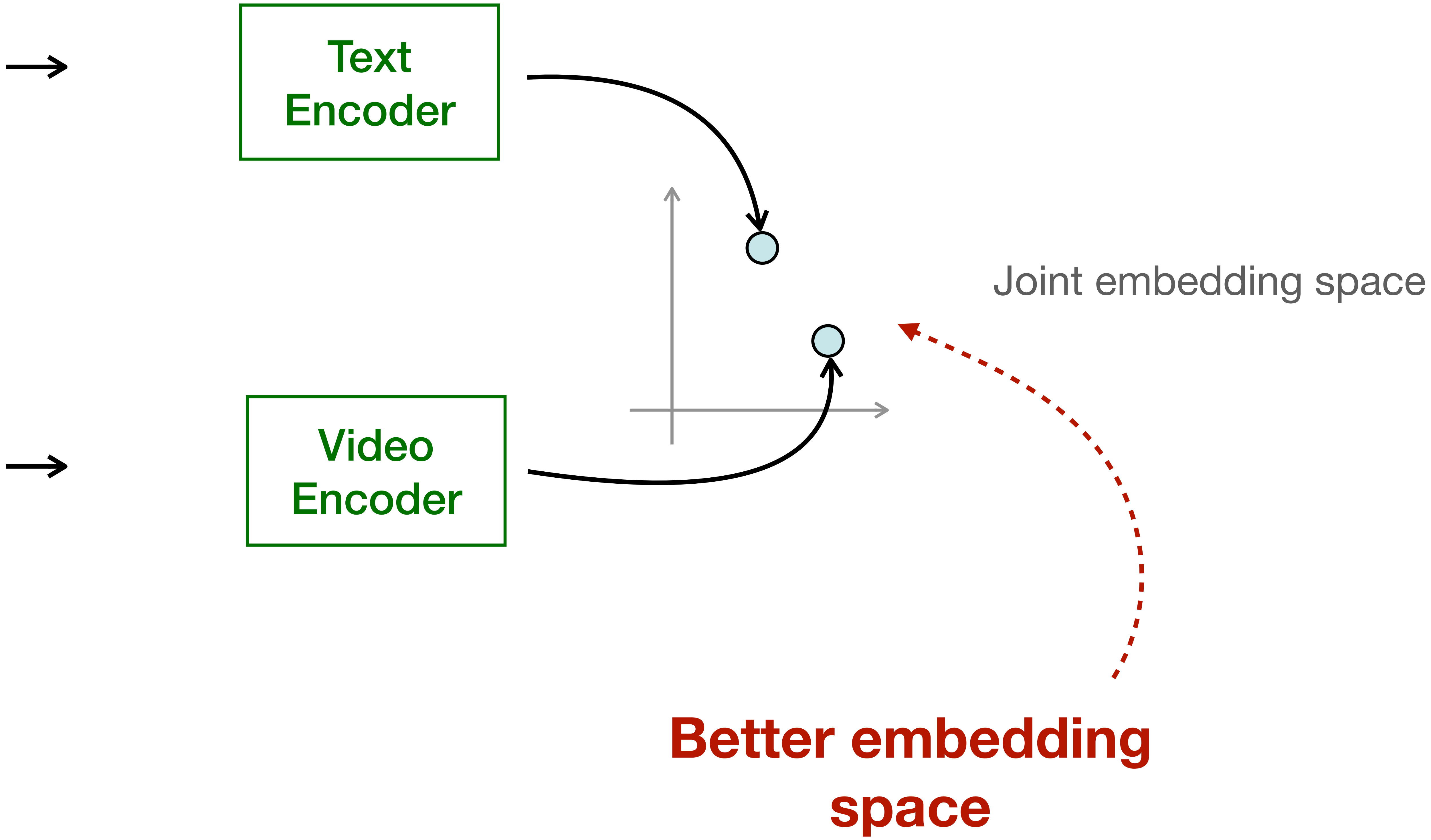
Caption: Matt Swanson gives a tip to use buckets to direct the path of the ball

ASR: so it's not going to really show

Caption: Making a bow with two colors

HowToCaption Method

Speaker prepares demonstration with the dog and toothpaste



HowToCaption Results

Video-Text Training Data	YouCook2 R10↑	YouCook2 MR↓	MSR-VTT R10↑	MSR-VTT MR↓
- (zero-shot)	23.6	69	70.6	3
HowTo100M with ASRs	39.3	20	61.7	5
HowTo100M with dist. sup.	30.3	34	66.3	5
HTM-AA (auto-aligned)	43.5	**15**	64.3	4
HowToCaption (ours)	**44.1**	**15**	**73.3**	**3**
VideoCC3M	21.7	84	67.1	4
WebVid2M	29.0	46	71.9	**3**

HowToCaption — Contributions

- Framework to obtain a **large-scale high-quality text-video dataset**
  - **No human supervision needed**
  - Only noisy ASR as input
  - Aligning&Filtering improves the quality even further

- **HowToCaption-dataset**
  - 25M aligned text-video pairs
  - human-style captions

What would be next unconventional way to leverage LLMs?

Thanks!