Attention (as Discrete-Time Markov) Chains

TEL AVIV UNIVERSITY · SIC Saarland Informatics Campus · MAX PLANCK INSTITUTE FOR INFORMATICS

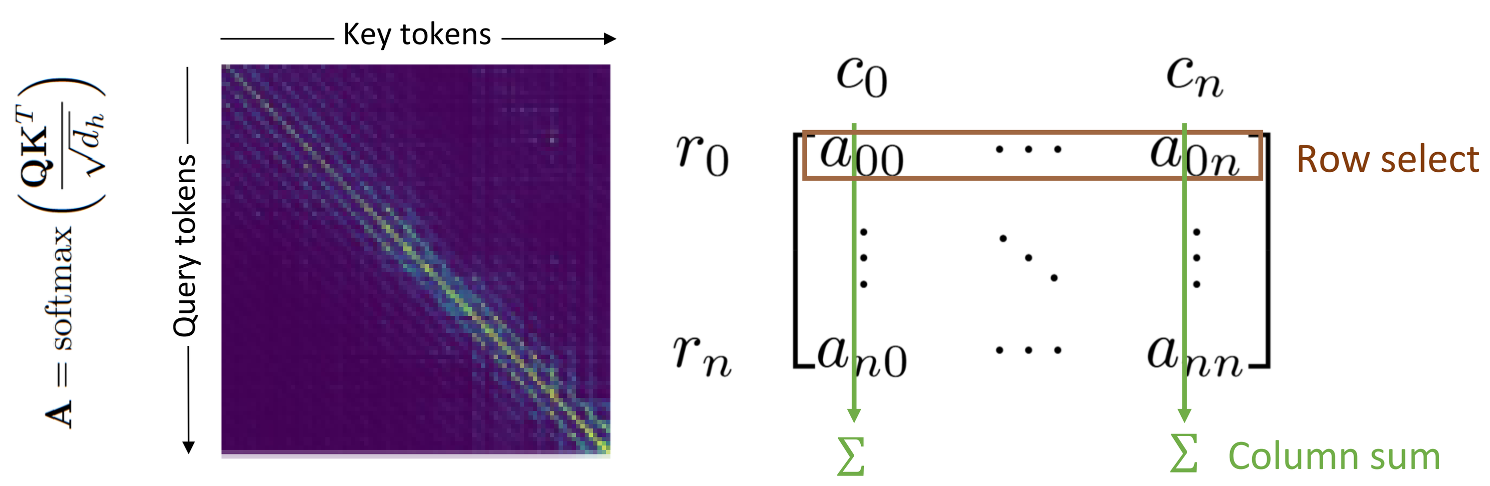
Yotam Erel*, Olaf Dünkel*, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, Amit H. Bermano

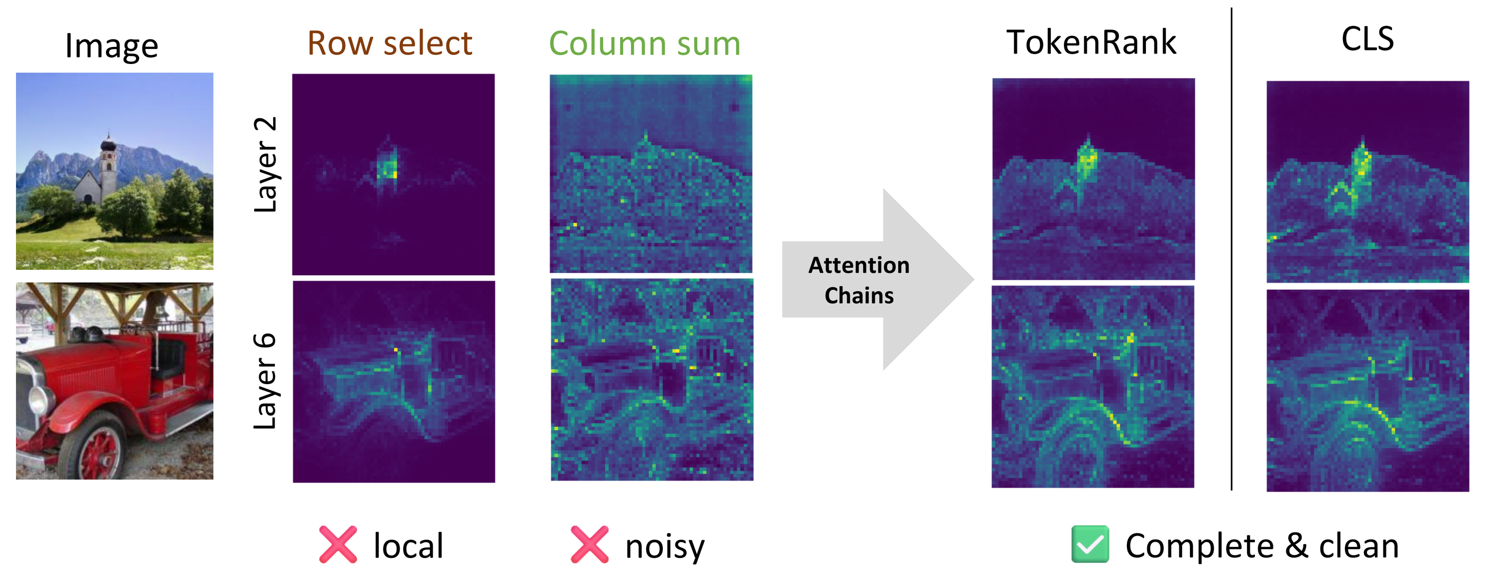
*Denotes equal contribution

Project Page

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation: How to visualize attention matrices?

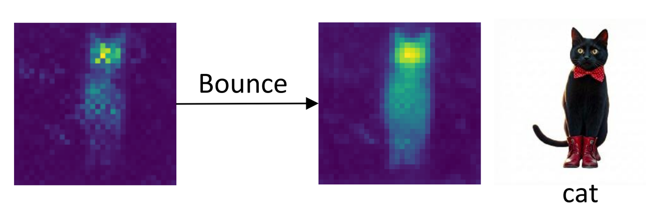




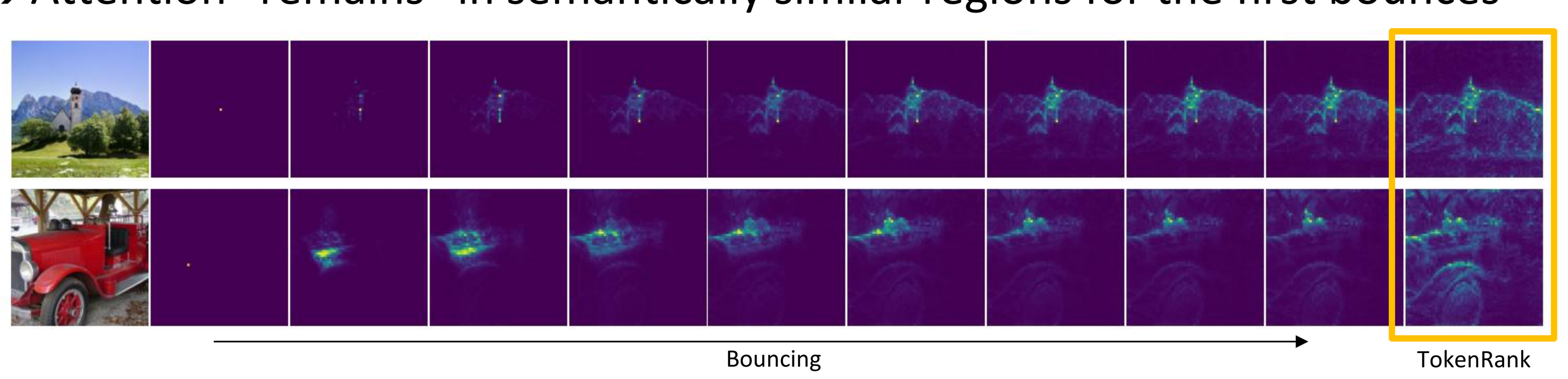
Metastable states of Markov Chain

→Regions where chain remains longer

→Bouncing consolidates object information



→Attention "remains" in semantically similar regions for the first bounces



Eigenvalues of Markov Chain

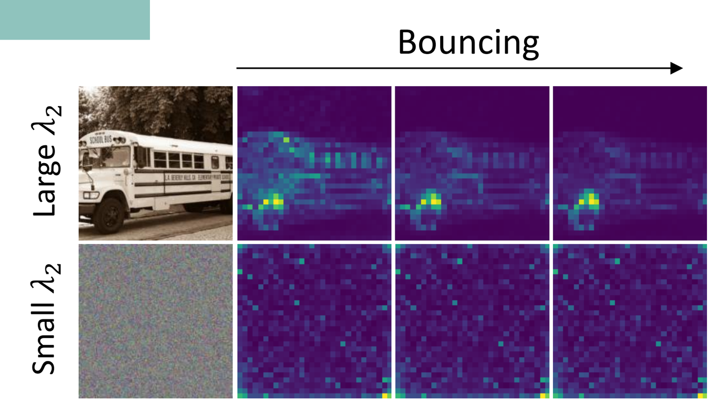
1st eigenvalue ($\lambda_1 = 1$): steady state

2nd eigenvalue (λ_2)

→ Tied to the convergence rate

$$\lambda_2 \sim \frac{1}{\text{DTMC convergence rate}}$$

→ Noisy attention heads have smaller λ_2



References

[1] Lawrence Page, et al. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

[2] Susung Hong et al. Improving sample quality of diffusion models using self-attention guidance. In ICCV, 2023.

[3] Junjiao Tian et al. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In CVPR, 2024.

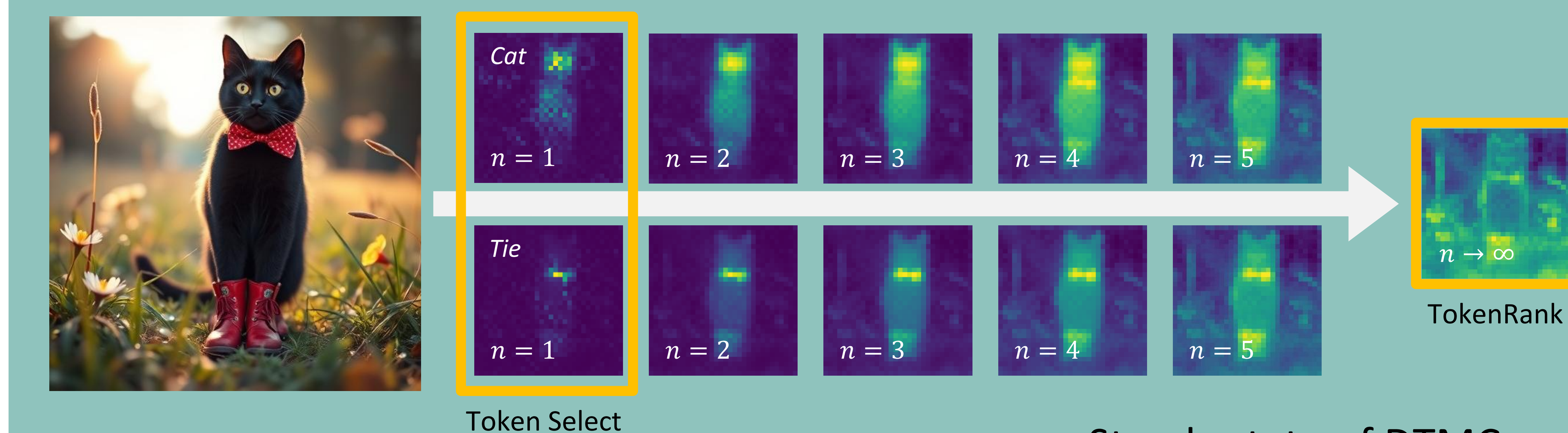
[4] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. In ICML, 2025.

TL;DR: TokenRank - PageRank for transformers

Attention as a DTMC

States of the Markov chain ⟶ **Tokens**

Prob. transitions to other states ⟶ **Attention weights**



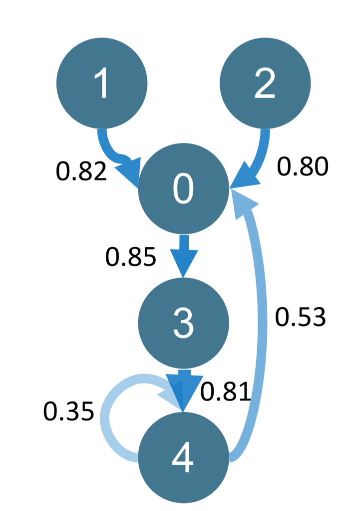
Steady state of DTMC

Bouncing along DTMC: $v_{n+1}^T = v_n^T \cdot \mathbf{A}$ $\quad n \to \infty$: $v_{SS}^T = v_{SS}^T \mathbf{A}$

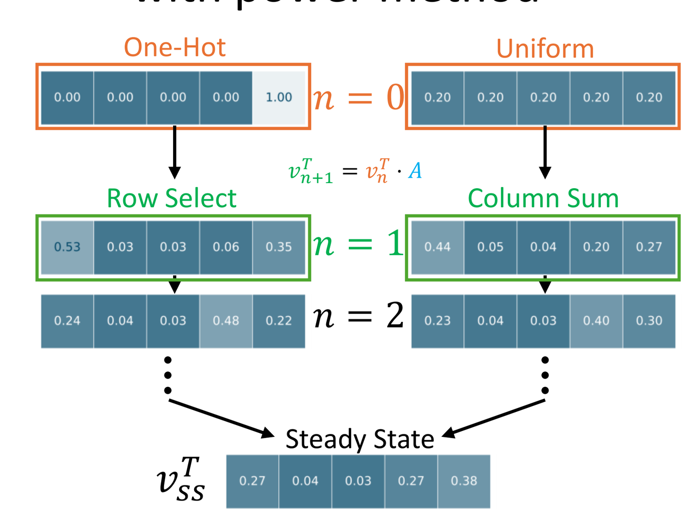
Attention Matrix **A**

	0	1	2	3	4
0	0.01	0.04	0.02	0.85	0.08
1	0.82	0.04	0.07	0.01	0.06
2	0.80	0.07	0.03	0.05	0.05
3	0.06	0.06	0.03	0.04	0.81
4	0.53	0.03	0.03	0.06	0.35

Markov Process



Attention Propagation with power method

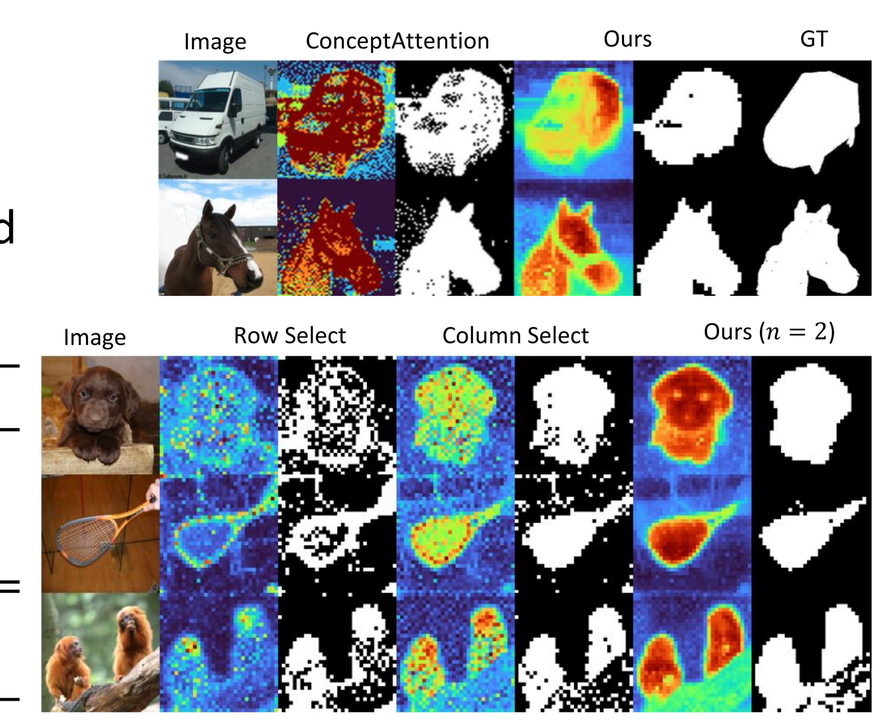


Zero-shot semantic segmentation with **multi-bounce attention**

Row Select: <text token> attends to image tokens

Column Select: Image tokens attend to <text token>

Ours: Attention bouncing via power method



Method	Architecture	Acc ↑	mIoU ↑	mAP ↑
FLUX row-select	FLUX DiT	73.96	54.65	82.64
FLUX column-select	FLUX DiT	80.55	64.02	87.20
Concept Attention	FLUX DiT	83.07	**71.04**	90.45
Ours w/o λ_2	FLUX DiT	<u>84.00</u>	70.02	<u>94.28</u>
Ours	FLUX DiT	**84.12**	<u>70.20</u>	**94.29**

→ Consolidation of semantic object maps via attention bouncing

→ λ_2-weighted head averaging improves results by favoring less noisy heads

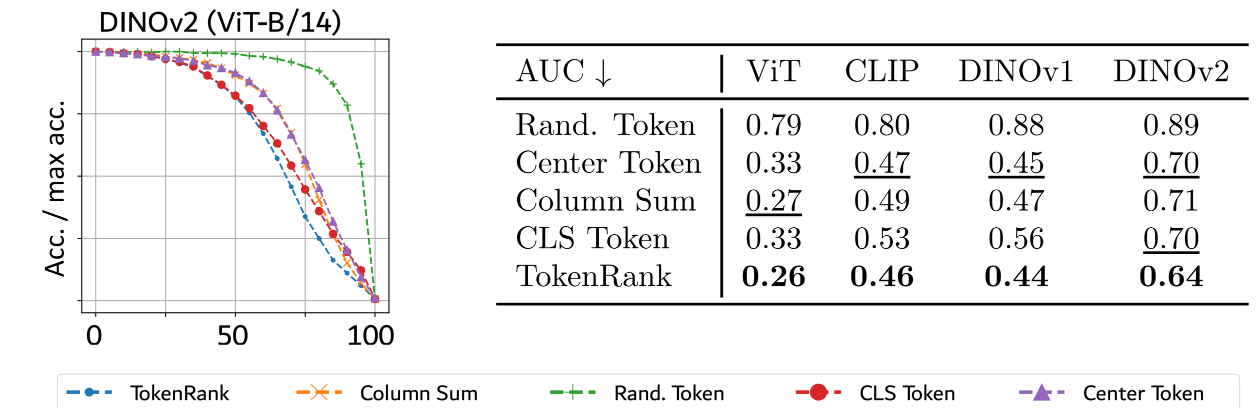
→ SOTA for zero-shot semantic segmentation

Experiments with TokenRank

Relevance of global token importance

TokenRank for downstream tasks

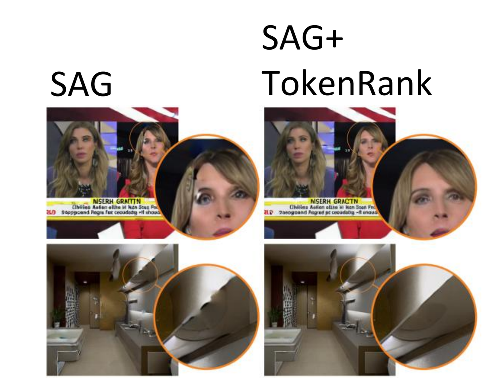
I) Masking Most Influential Tokens



AUC ↓	ViT	CLIP	DINOv1	DINOv2
Rand. Token	0.79	0.80	0.88	0.89
Center Token	0.33	<u>0.47</u>	<u>0.45</u>	<u>0.70</u>
Column Sum	<u>0.27</u>	0.49	0.47	0.71
CLS Token	0.33	0.53	0.56	<u>0.70</u>
TokenRank	**0.26**	**0.46**	**0.44**	**0.64**

→ Larger Classification drops with TokenRank

II) Self-Attention Guidance with TokenRank



Quantitative Results

Method	IS ↑	FID ↓	KID ↓
SD1.5	16.32	45.77	0.018
SAG	17.69	52.48	0.023
SAG+TokenRank	**18.37**	**50.14**	**0.021**

→ Higher generation quality when refining tokens for important features ranked with TokenRank

III) DiffSeg with TokenRank anchor sampling

Semantic segmentation on COCO-stuff

Method	mACC ↑	mIoU ↑
Uniform Grid	72.50	43.60
TokenRank Grid	**84.97**	**44.87**

→ Anchor sampling based on TokenRank improves DiffSeg

→ More accurate ranking of token importance improves various downstream tasks

Takeaway

Multi-bounce attention improves downstream tasks via **attention consolidation** and better **global token importance**