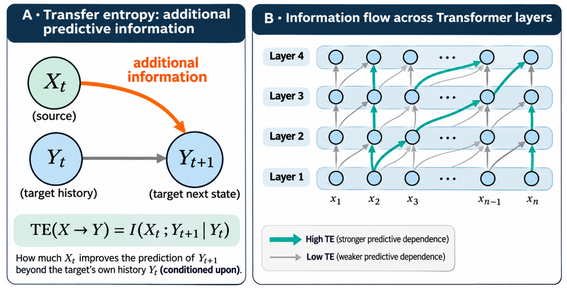
Transfer Entropy as a Measure of Information Flow in VLMs and LLMs

Jessica E. Liang[1,2], Jianbo Shi[1]
[1]Department of Computer and Information Science, University of Pennsylvania
[2]Operations Research Center, Massachusetts Institute of Technology

Introduction and Motivation

Transformers drive VLMs and LLMs, yet the roles of individual layers, tokens, and heads remain opaque. Attention weights reveal where probability mass is placed, while static similarity cannot resolve the direction of information transfer. We propose transfer entropy (TE) as a directed-information diagnostic for interpretation and pruning.

Transfer Entropy for Transformer Layers



A Tractable TE Approximation

Exact conditional-information estimation is intractable for large hidden states. With local Gaussian perturbations, linearized Jacobian responses, and a rank-one covariance approximation, we derive

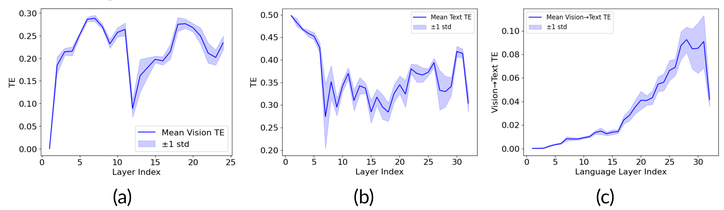
$$\widehat{\mathrm{TE}}_\ell \approx \frac{1}{2(B-1)} \sum_{i=1}^{B-1} \cos^2(\Delta z_{\ell+1,i}, \Delta z_{\ell,i}) \,.$$

- Cache activations from a fixed probe batch around an optimizer step.
- Form neighboring-item differences $\Delta z_{\ell,i} = z_{\ell,i+1} - z_{\ell,i}$.
- Average ridge-stabilized squared cosine values across pairs and steps.

Turning TE into Structured Pruning

- Layers: high averaged TE_ℓ indicates that layer $\ell+1$ may add limited unique information.
- Tokens: prune large positive gaps $\mathrm{TE}_{\mathrm{out}} - \mathrm{TE}_{\mathrm{in}}$, the broadcaster tokens.
- Heads: combine TE with CLS/EOT routing patterns and prune *low-TE benign* heads.

LLaVA-1.5-7B: Where Does Vision Meet Text?



(a) (b) (c)

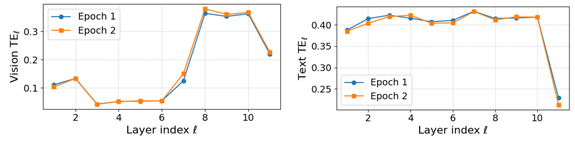
Figure 1. Statistics of information flow in LLaVA-1.5-7B. Mean and standard deviation of the TE across 5-epoch LoRA (rank 4) fine-tuning on 2 datasets (MSCOCO and Flickr8k). (a) vision tower, (b) text tower, (c) vision→text.

LLaVA: TE-Guided Decoder-Layer Pruning

Table 1. Decoder-layer pruning accuracy on LLaVA-1.5-7B (VQAv2). The unpruned model achieves 0.6480.

# Layers Pruned	TE Pruning	Streamline Pruning	Short-LVLM
2	0.5370	0.4498	0.4548
4	0.3798	0.2937	0.2832
6	0.2867	0.1352	0.2248
8	0.2737	0.1220	0.1472

CLIP ViT-B/16: Information Flows and Layer Pruning



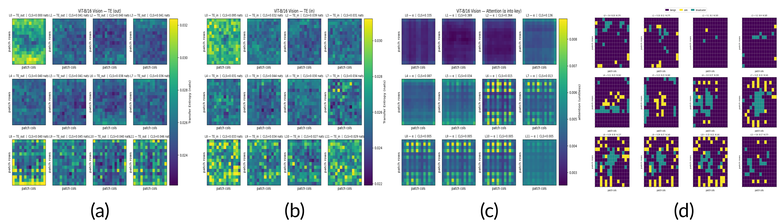
(a) (b)

Figure 2. Layerwise TE profile of CLIP ViT-B/16 in LoRA ($r = 8$) fine tuning. (a) Vision TE_ℓ. (b) Text TE_ℓ.

Table 2. Retrieval performance (MSCOCO) of CLIP ViT-B/16 under different pruning strategies.

Method	Pruned Layers	Image-to-Text Recall (%)			Text-to-Image Recall (%)		
		R@1	R@5	R@10	R@1	R@5	R@10
No Pruning	–	42.20	66.21	74.94	26.66	48.94	58.91
Our TE pruning	2V+2T	27.36	51.44	62.78	16.67	36.50	47.03
	4V+4T	15.95	34.81	45.51	9.53	24.08	33.11
	5V+5T	11.79	28.42	38.68	7.07	19.49	27.92
Short-LVLM	2V+2T	25.65	49.34	60.63	15.91	35.16	45.65
	4V+4T	13.10	30.33	40.28	7.95	20.85	29.29
	5V+5T	8.02	20.62	28.81	4.87	14.19	20.91
Streamline pruning	2V+2T	26.83	50.26	60.83	16.18	35.38	45.66
	4V+4T	12.02	28.49	38.18	7.57	20.03	28.37
	5V+5T	9.34	23.07	31.63	5.67	16.15	23.54

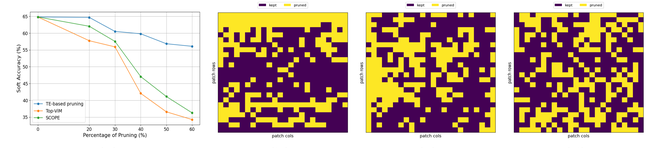
Information Flow in and out of Tokens in VLMs



(a) (b) (c) (d)

Figure 3. CLIP ViT-B/16 (vision) tokens from the 14×14 patches. (a) TE_out, (b) TE_in, (c) average attention, (d) classification of vision token roles.

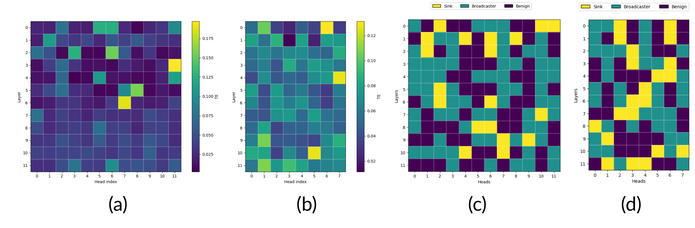
LLaVA: Vision-Token Pruning on VQAv2



(a) (b) (c) (d)

Figure 4. (a) VQAv2 soft accuracy versus the percentage of pruned LLaVA-1.5-7B vision tokens. (b)(c)(d) pruning 40% of LLaVA-1.5-7B vision tokens on VQAv2. (b) TE-guided broadcaster pruning, (c) Top-ViM, and (d) SCOPE.

CLIP: Attention Heads Classification and Pruning on MSCOCO

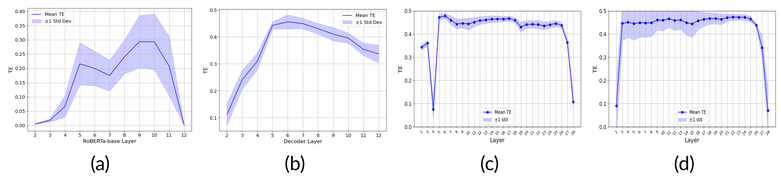


(a) (b) (c) (d)

Figure 5. TE of each attention head in every layer for CLIP ViT-B/16: (a) Vision tower, (b) text tower. Attention heads classification in each layer: (c) Vision tower, (d) Text tower.

Pruning per Layer	Method	I→T R@1	I→T R@5	I→T R@10	T→I R@1	T→I R@5	T→I R@10
None	No Pruning	42.20	66.21	74.94	26.66	48.94	58.91
3V + 2T	TE	36.97	61.94	71.82	23.49	45.40	55.67
3V + 2T	SmartTrim	34.92	59.79	69.78	21.78	43.12	53.34
5V + 3T	TE	32.41	56.64	66.89	20.40	41.41	51.77
5V + 3T	SmartTrim	30.53	54.42	64.87	19.16	39.48	49.66

Information Flow in LLMs



(a) (b) (c) (d)

Figure 6. Statistics of information flow in LLMs on 8 datasets in GLUE benchmark. (a) RoBERTa-base, (b) T5 decoder, (c) Qwen-2.5-7B, (d) Llama-3.2-3B.

Conclusion

- TE reveals where information flows through VLMs and LLMs.
- The same information identifies redundant layers, token communication roles, and removable benign heads.
- TE-guided structured pruning better preserves predictive performance than the tested attention/saliency baselines at matched budgets.

The Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS), 2026