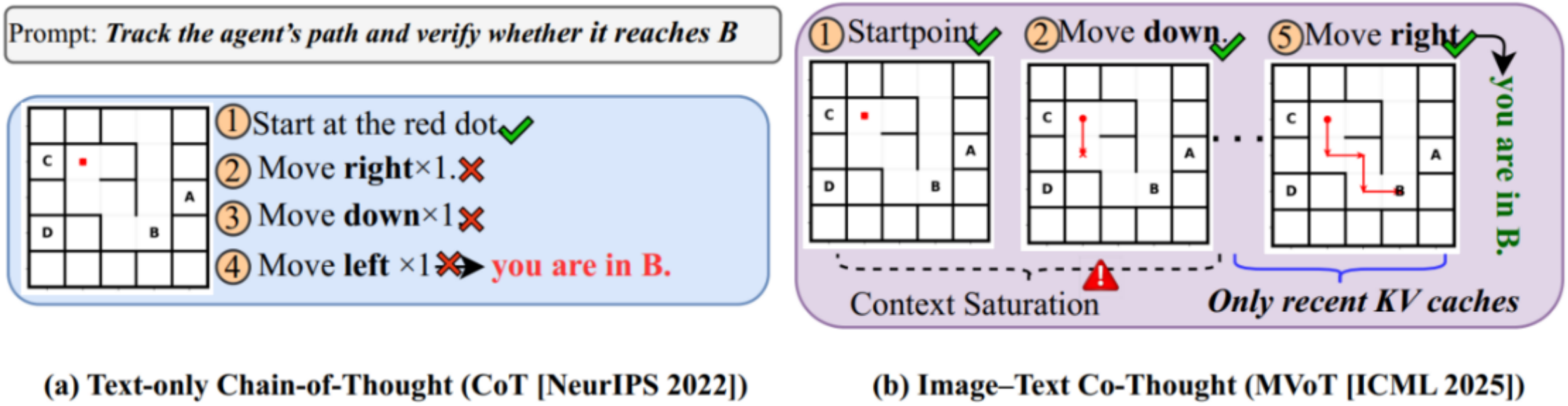
Efficient Multimodal Spatial Reasoning via Dynamic and Asymmetric Routing

Goal: Achieve strong adaptation and reasoning performance with **minimal additional parameters and system overhead**

Yixian Shen

Multimodal Tokens Saturate the Context Window

- ❑ Text-only reasoning is often insufficient for spatial tasks
- ❑ Multimodal Chain-of-Thought improves spatial understanding
- ❑ But interleaved text–image tokens rapidly saturate the context window
- ❑ Context overflow degrades long-horizon reasoning



(a) Text-only Chain-of-Thought (CoT [NeurIPS 2022])

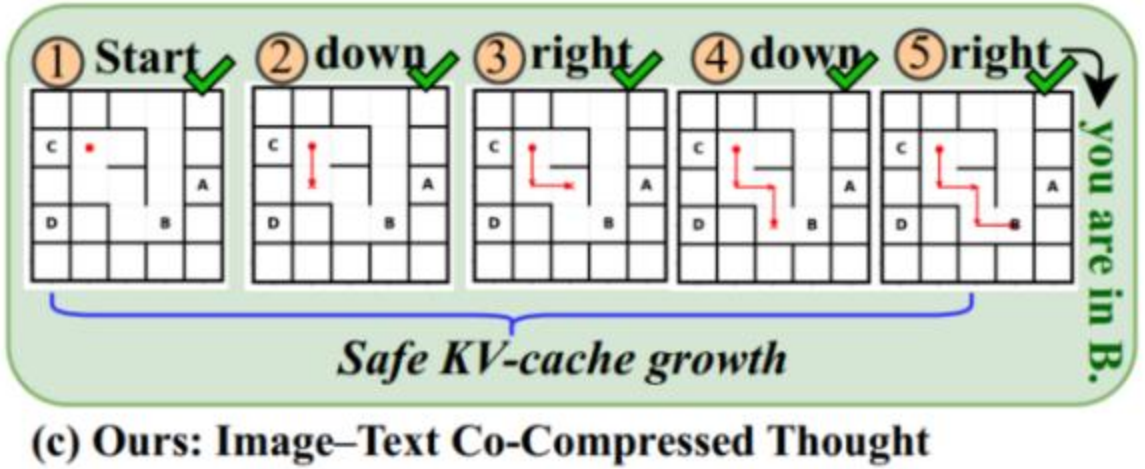
(b) Image–Text Co-Thought (MVoT [ICML 2025])

Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

Li, Chengzu, et al. "Imagine while reasoning in space: Multimodal visualization-of-thought." arXiv preprint arXiv:2501.07542 (2025).

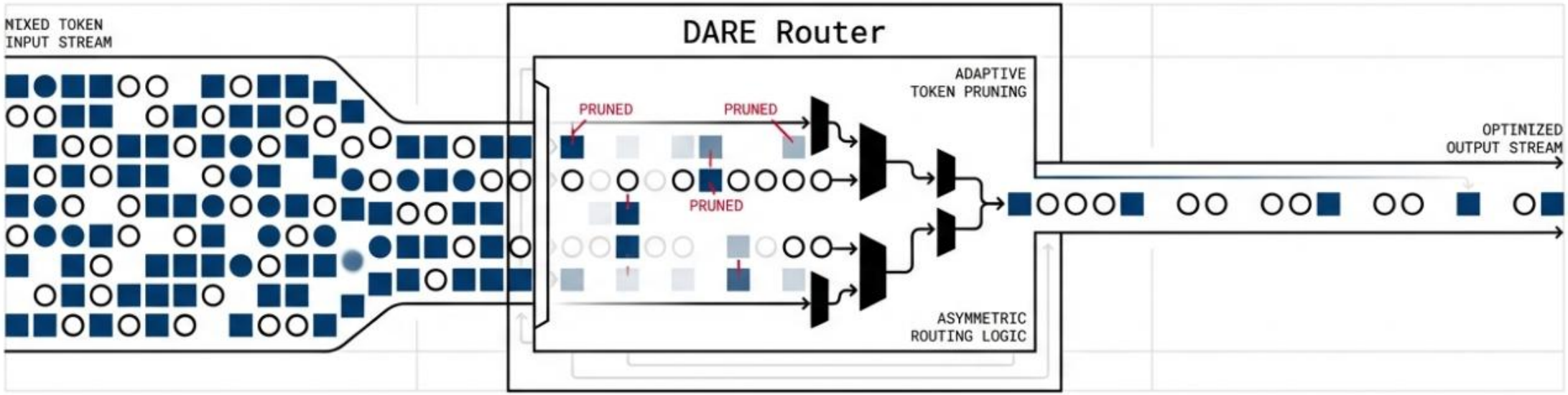
DARE: Efficient Multimodal Spatial Reasoning via *Dynamic* and *Asymmetric* Multimodal Token Routing

❑ Intra- and Inter-Hop Adaptive Token Compression for MLLMs



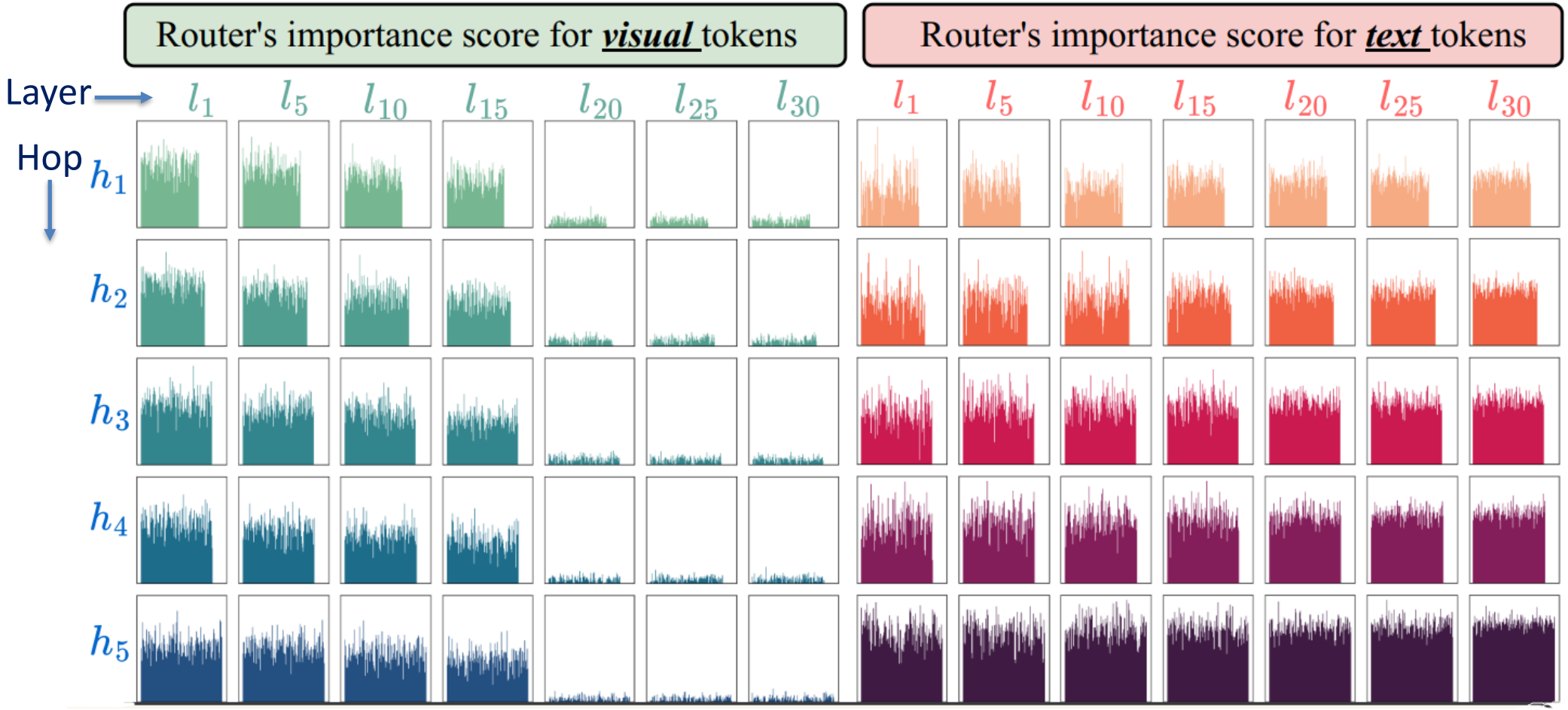
(c) Ours: Image–Text Co-Compressed Thought

- ➢ End-to-end differentiable routing for text and vision tokens
- ➢ Fusion-aware routing enables aggressive yet safe pruning
- ➢ Safe KV-cache growth for long-horizon reasoning



Asymmetric Redundancy in Multimodal Reasoning

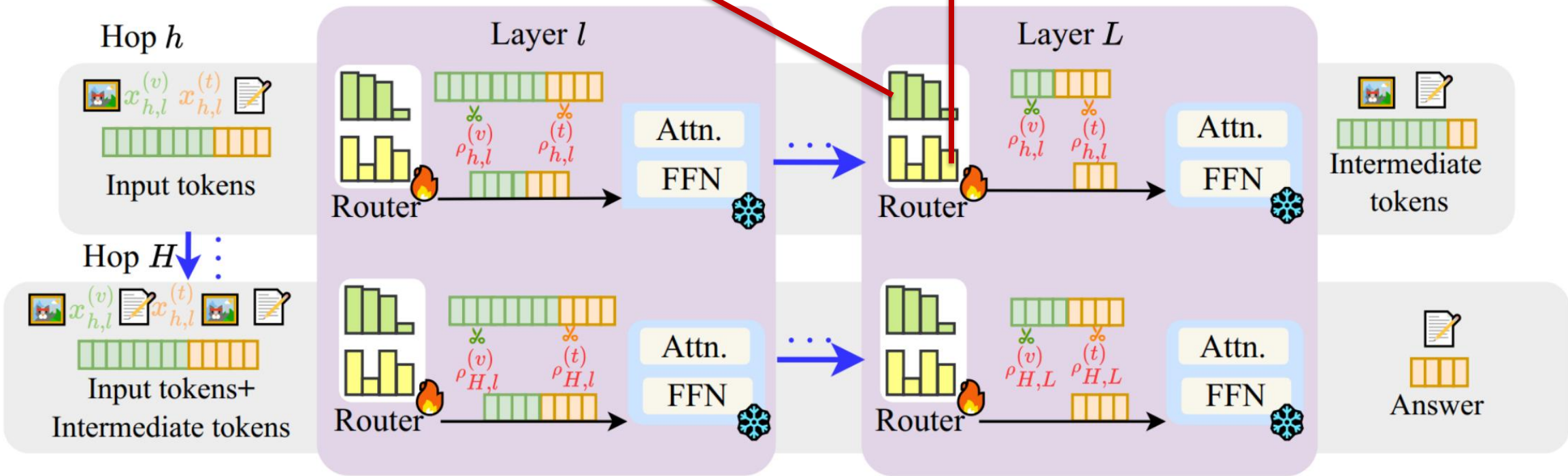
- ❑ Text and vision tokens exhibit different redundancy patterns
- ❑ Routing importance evolves differently across reasoning steps
- ❑ Motivates modality-specific routing strategies



DARE: A Fully Differentiable and Dynamic Routing Framework

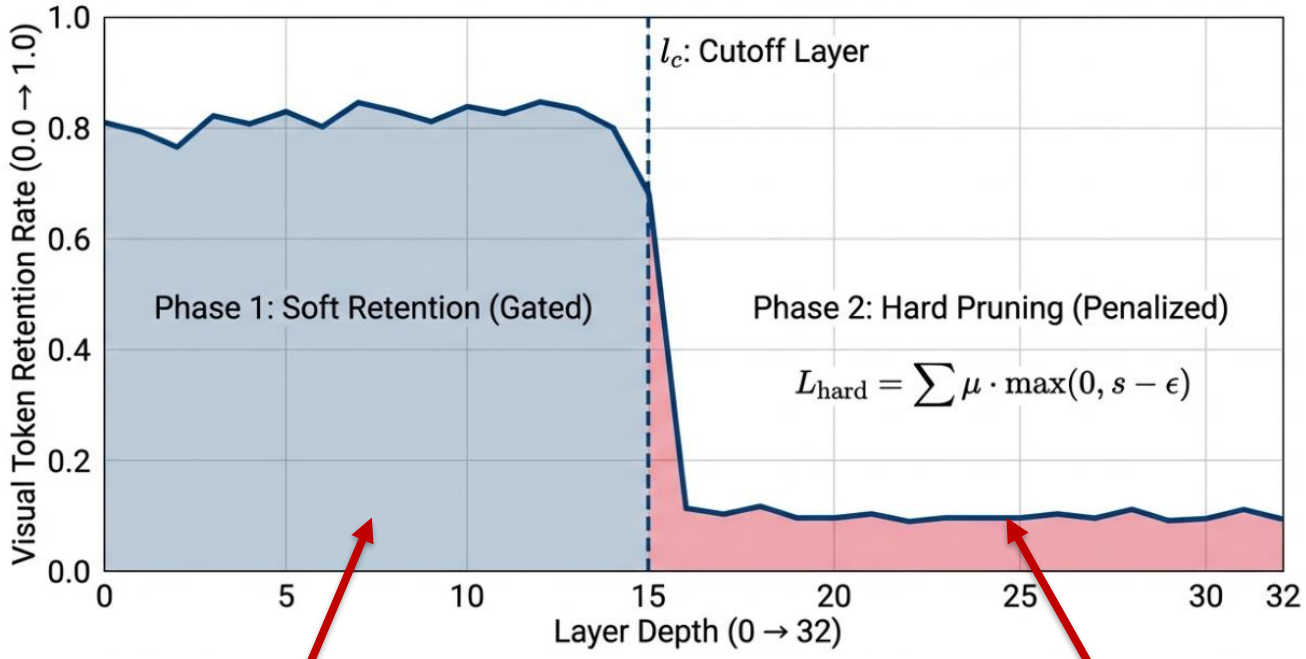
- ❑ **Modality-specific, learnable routing for text and vision tokens**
- ❑ **End-to-end differentiable routing objectives**
- ❑ **Adaptive token compression across layers and hops**

$$\mathcal{L}_{\text{ratio}}^{(\text{v})} = \frac{1}{HL}\sum_{h=1}^{H}\sum_{l=1}^{L}\left(\hat{\rho}_{h,l}^{(\text{v})} - \rho_{\text{target}}^{(\text{v})}\right)^2 \qquad \mathcal{L}_{\text{ratio}}^{(\text{t})} = \frac{1}{HL}\sum_{h=1}^{H}\sum_{l=1}^{L}\left(\hat{\rho}_{h,l}^{(\text{t})} - \rho_{\text{target}}^{(\text{t})}\right)^2$$



The Asymmetric Redundancy

- ❑ Information cliff in visual token utility
- ❑ Greedy cutoff identifies where vision tokens become redundant
- ❑ Aggressive pruning applied after the cutoff layer

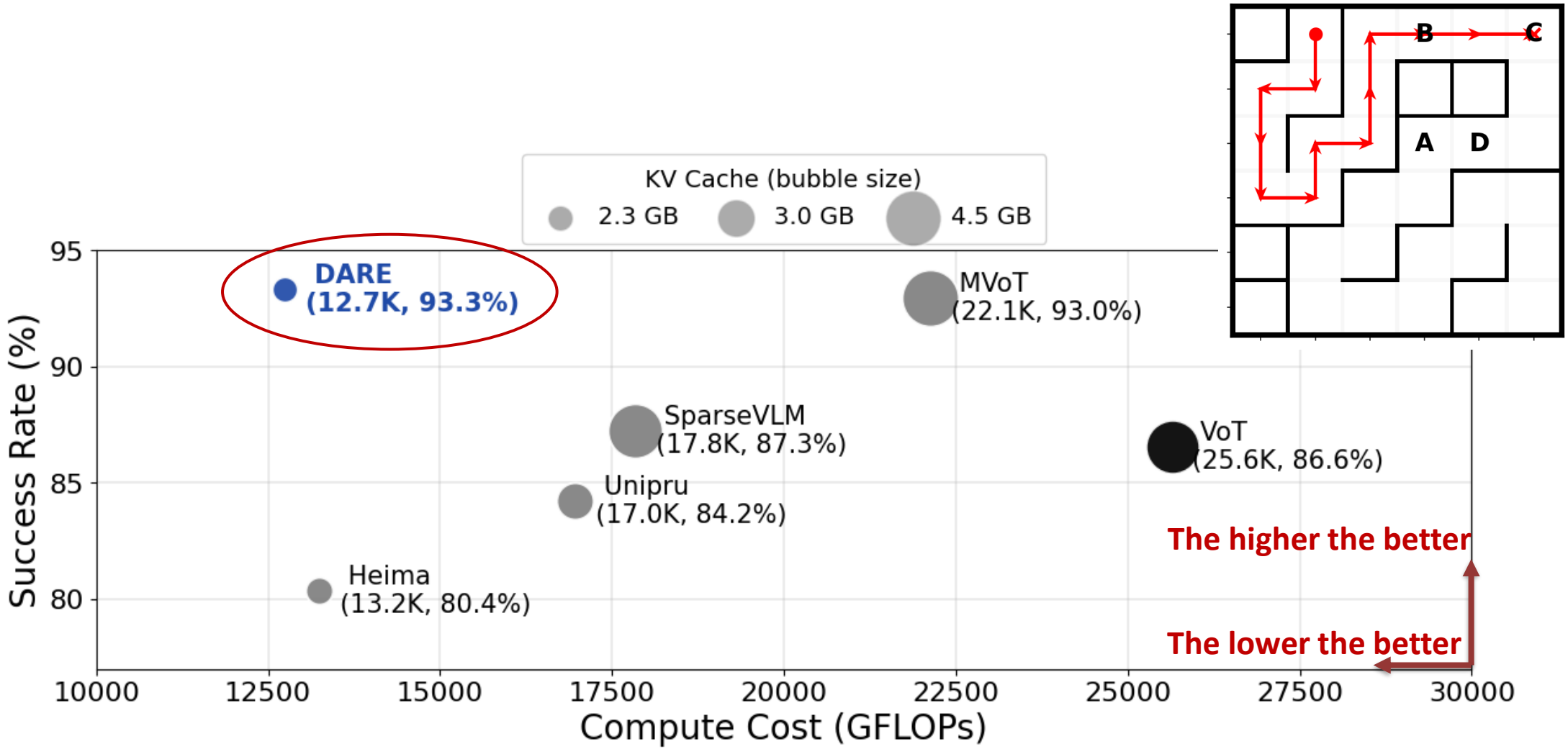


$$\mathcal{L}_{\text{ratio}}^{(\text{v})} = \frac{1}{HL} \sum_{h=1}^{H} \sum_{l=1}^{L} \left(\hat{\rho}_{h,l}^{(\text{v})} - \rho_{\text{target}}^{(\text{v})} \right)^2$$

$$\mathcal{L}_{\text{hard}}^{(\text{v})} = \sum_{h=1}^{H} \sum_{l=l_c+1}^{L} \sum_{i=1}^{N_{h,l}^{(\text{v})}} \mu \cdot \max(0, s_{h,l}^{(i,\text{v})} - \epsilon)$$

Dynamic Spatial Reasoning---MAZE

❑ **DARE achieves the highest success rate with the lowest compute and KV-cache cost.**



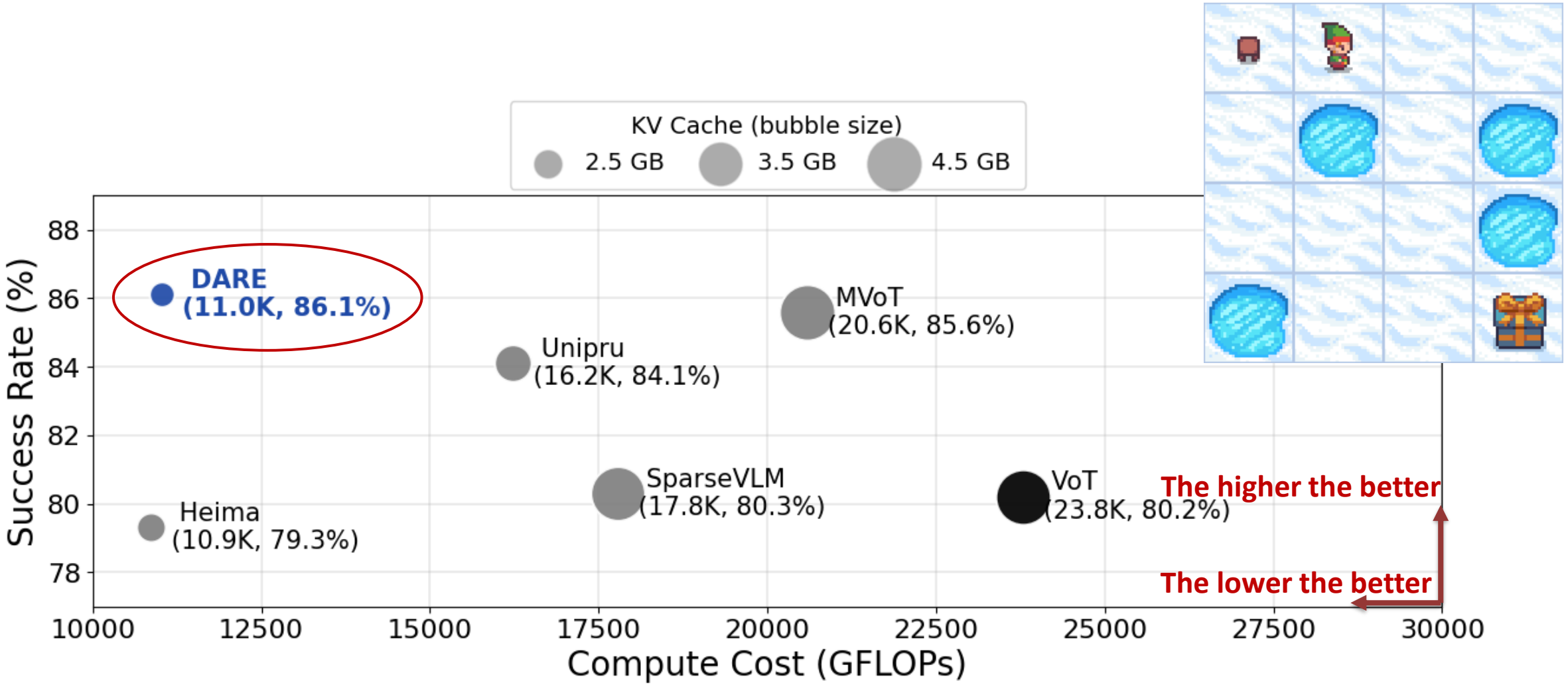
Zhang, Yuan, et al. "Sparsevlm: Visual token sparsification for efficient vision-language model inference." *arXiv preprint arXiv:2410.04417* (2024).

Wu, Wenshan, et al. "Mind's eye of LLMs: visualization-of-thought elicits spatial reasoning in large language models." *Advances in Neural Information Processing Systems* 37 (2024): 90277-90317.

Shen, Xuan, et al. "Efficient reasoning with hidden thinking." *arXiv preprint arXiv:2501.19201* (2025).

Dynamic Spatial Reasoning---FrozenLake

❑ **DARE achieves the best success–efficiency trade-off on FrozenLake.**



Zhang, Jintian, et al. "Lightthinker: Thinking step-by-step compression." *arXiv preprint arXiv:2502.15589* (2025).

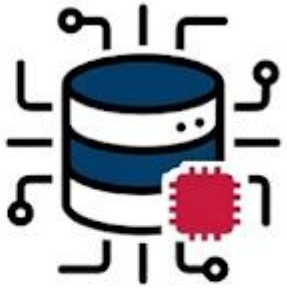
Summary: A Scalable Recipe for Multimodal Reasoning

Dynamic Routing

Intra- and Inter-hop adaptation prevents redundancy.

Asymmetric Compression

Exploits the visual-to-verbal information cliff.



System Efficiency

>40% Compute & Memory Savings.

❑ Limitations

- ➢ Requires tuning of text–vision compression targets
- ➢ Requires tuning of hard pruning thresholds