UPop: Unified and Progressive Pruning for Compressing Vision-Language Transformers

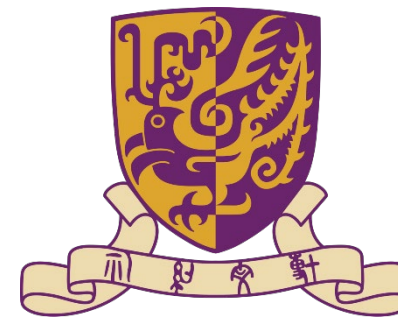Dachuan Shi[1,2], Chaofan Tao[3], Ying Jin[4], Zhendong Yang[1], Chun Yuan[1,✉] Jiaqi Wang[2,✉]
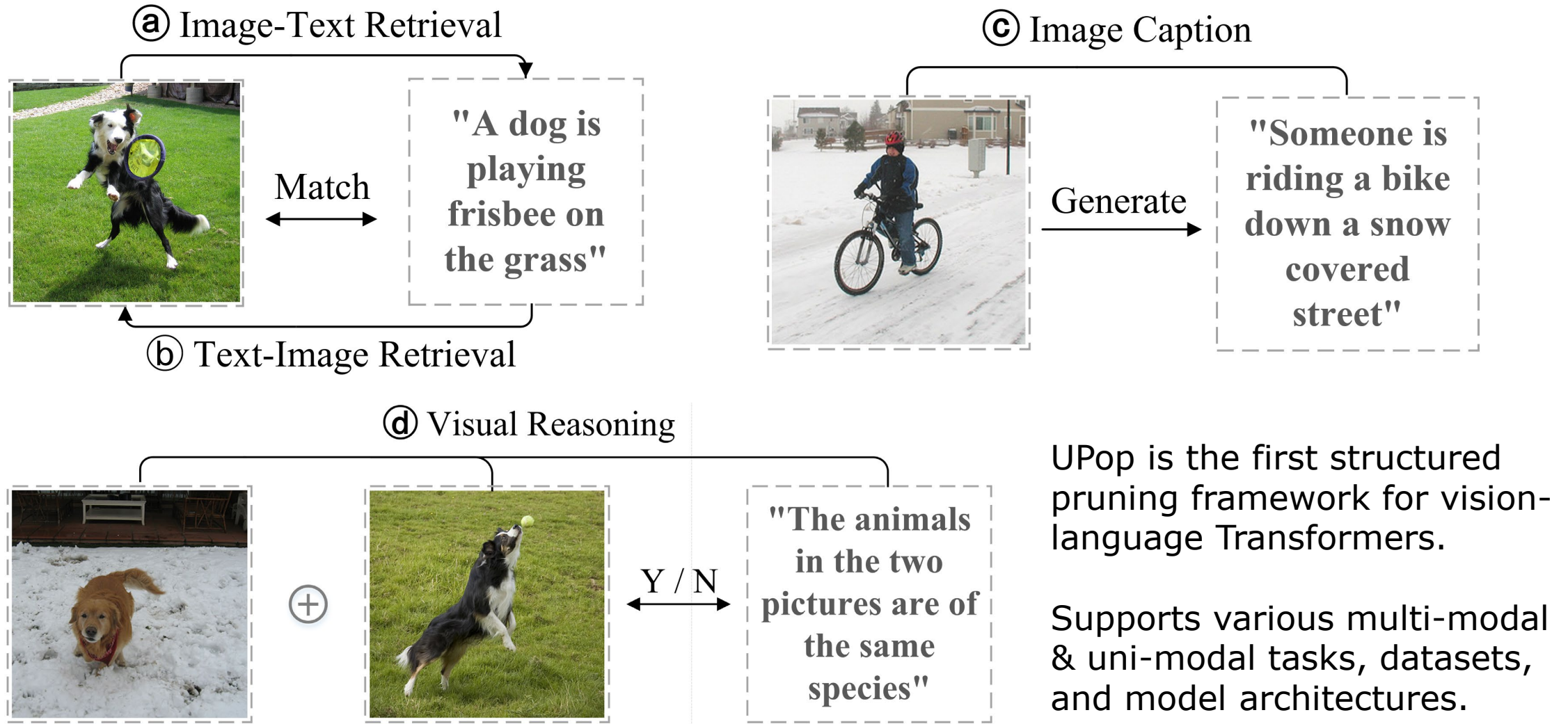
[1]Tsinghua University, [2]Shanghai AI Laboratory

[3]The University of Hong Kong, [4]The Chinese University of Hong Kong

Task: Compressing Vision-Language Transformers

ⓐ Image-Text Retrieval

Match

"A dog is playing frisbee on the grass"

ⓑ Text-Image Retrieval

ⓒ Image Caption

Generate

"Someone is riding a bike down a snow covered street"

ⓓ Visual Reasoning

⊕

Y / N

"The animals in the two pictures are of the same species"

UPop is the first structured pruning framework for vision-language Transformers.

Supports various multi-modal & uni-modal tasks, datasets, and model architectures.
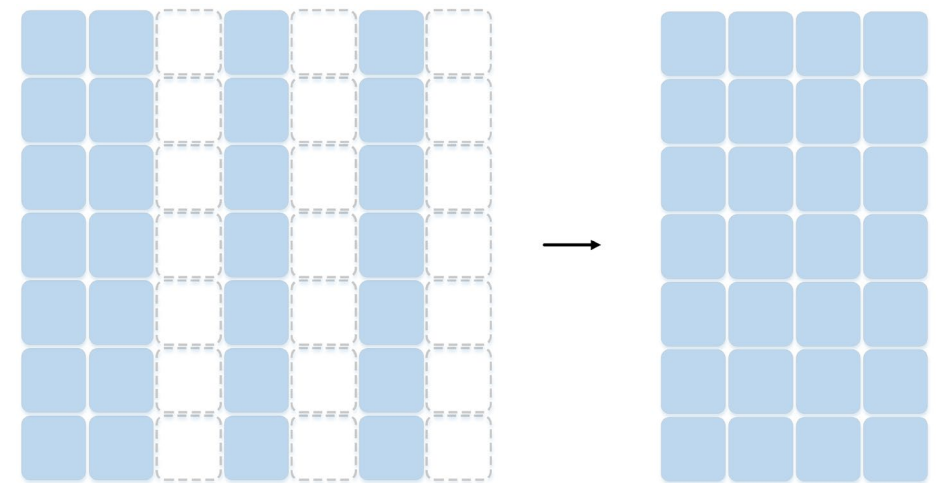
Task: Compressing Vision-Language Transformers

Vector-level structured
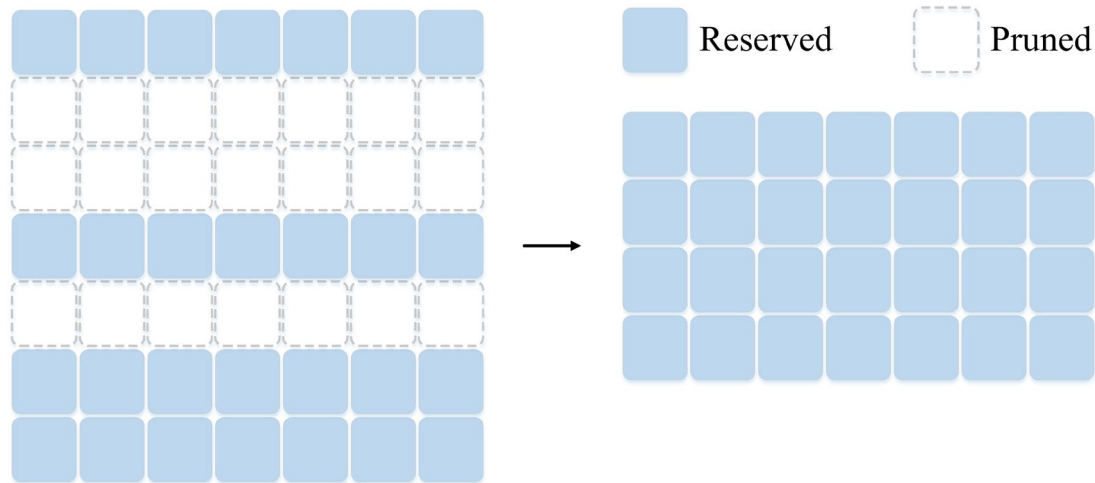
(a) Remove entire rows

(b) Remove entire columns

Motivation: Unified and Progressive Pruning

Mask-based Pruning (baseline, NLVR2: 76.4%)
- ○ Separately search, manual assignment, inefficient
- ○ Worse performance at the same compression ratio
- ○ Prune non-zero weights, hurt convergence
- ○ Support lower compression ratios
- ○ Partially structured

UPop (Ours, NLVR2: 81.1%)
- √ Unified search, adaptive assignment, efficient
- √ Better performance at the same compression ratio
- √ Prune zero weights, maintain convergence
- √ Support higher compression ratios
- √ Entirely structured

Question1: How to assign compression ratios?
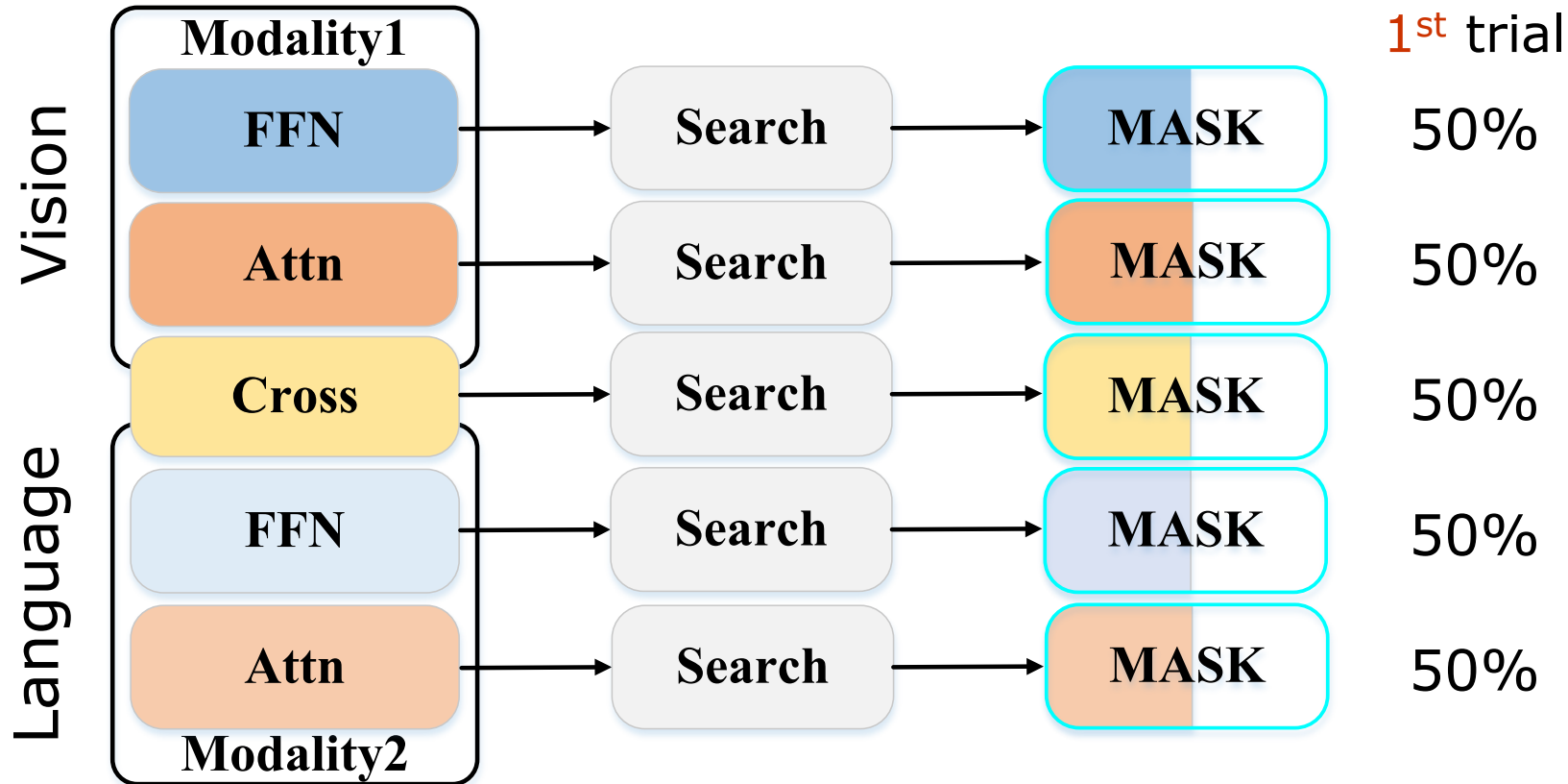Unified Search: From manually to adaptively.

Question2: How to prune weights from original models?
Progressive Pruning: Eliminate the weight gap.
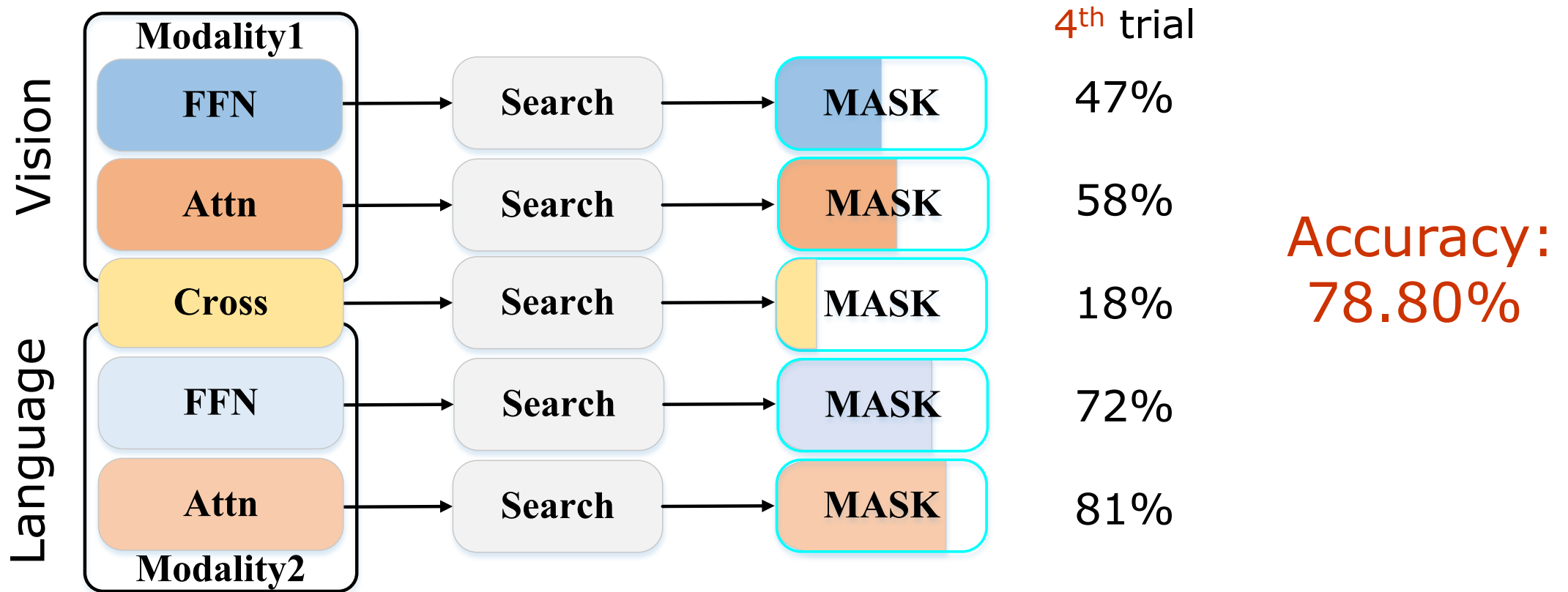
How to assign a 50% compression ratio in total ?

🤠 : So easy! Let's do a grid search!

1st trial

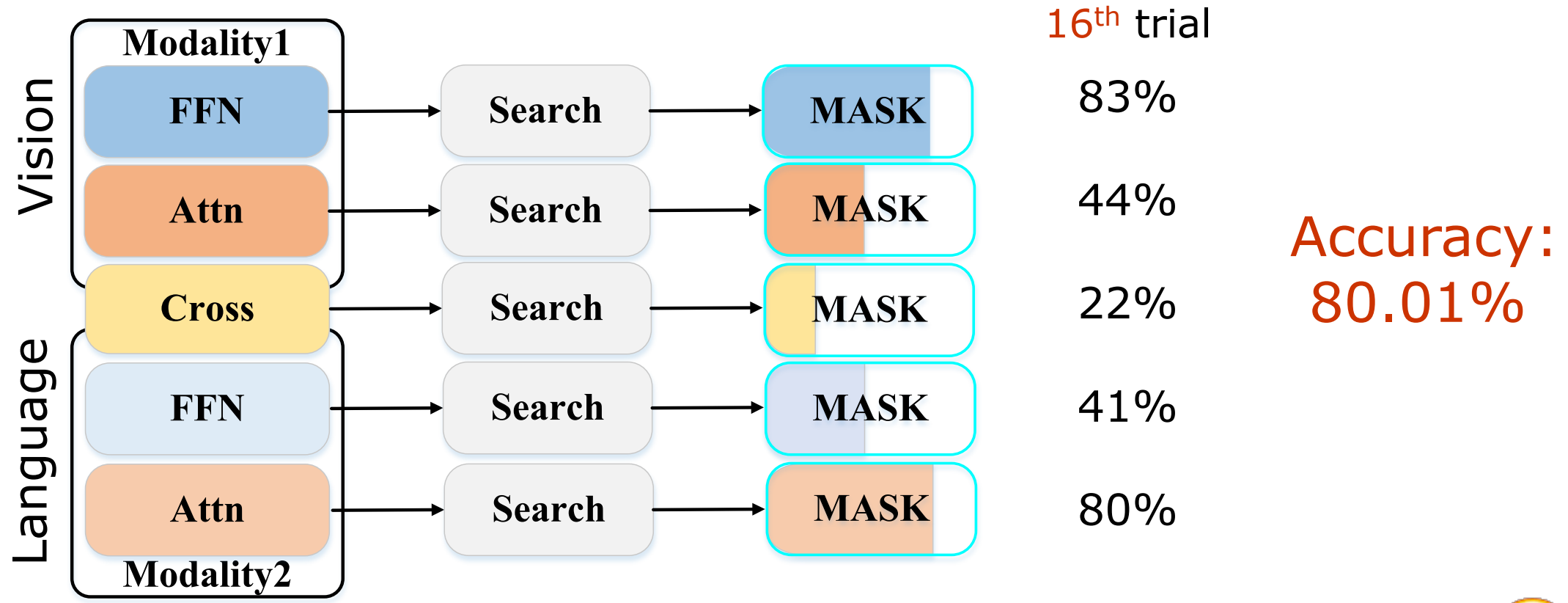| Modality1		
Vision	FFN → Search → MASK	50%
	Attn → Search → MASK	50%
	Cross → Search → MASK	50%
Language	FFN → Search → MASK	50%
	Attn → Search → MASK	50%

Accuracy: 76.44%

How to assign a 50% compression ratio in total ?

😐 : It's fine. We just need some patience.
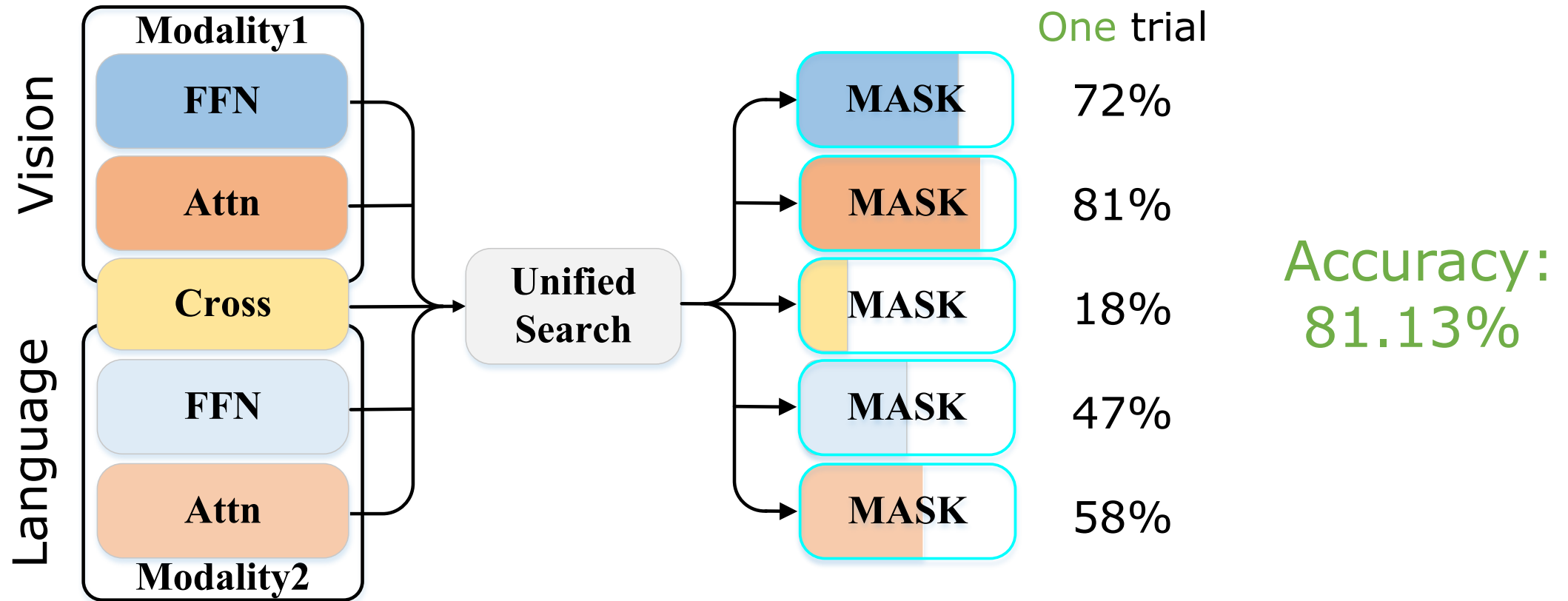
How to assign a 50% compression ratio in total ?

🌀 : So tedious. I am worn out.

16th trial

Vision	FFN	Search	MASK 83%
	Attn	Search	MASK 44%
	Cross	Search	MASK 22%
Language	FFN	Search	MASK 41%
	Attn	Search	MASK 80%

Accuracy: 80.01%

How about unified search? : 🤔

How to assign a 50% compression ratio in total ?

Unified search enables adaptive assignment !: 😎

One trial

Accuracy: 81.13%

🥹 : Rescued from the burden of repeated trials !

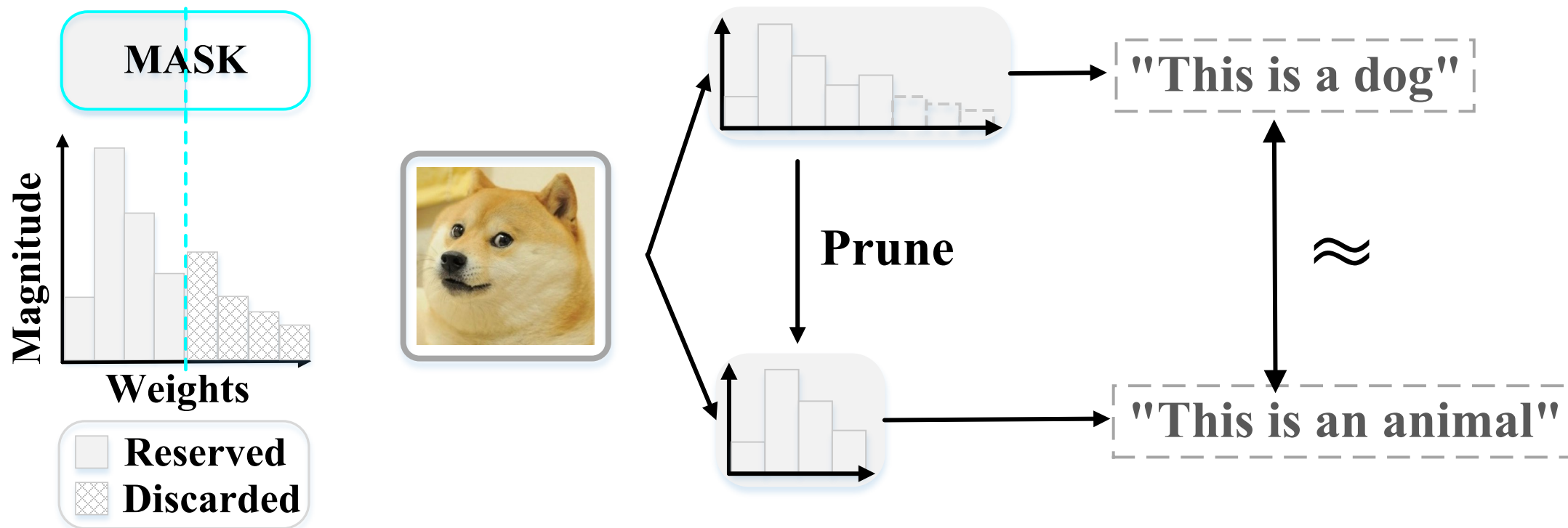# How to prune 50% weights from original models?

🤠 : It's simple! Prune all of them at once!
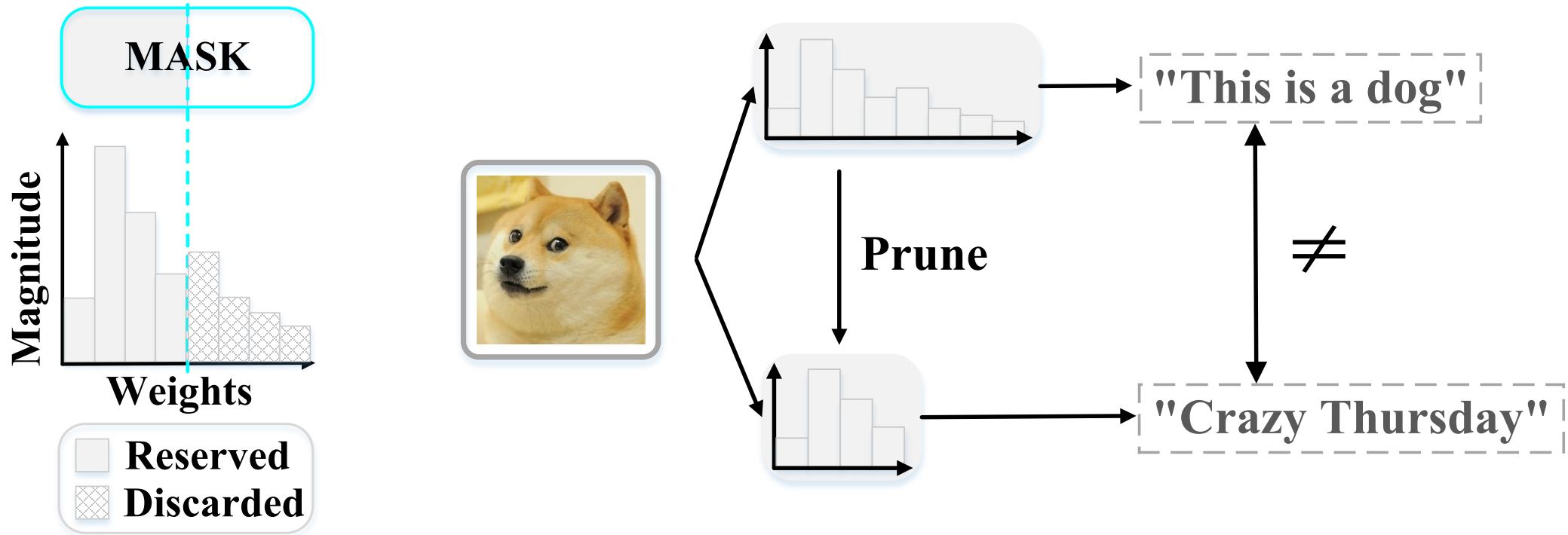
Weight gap between the searched model and pruned model ? : 🤔

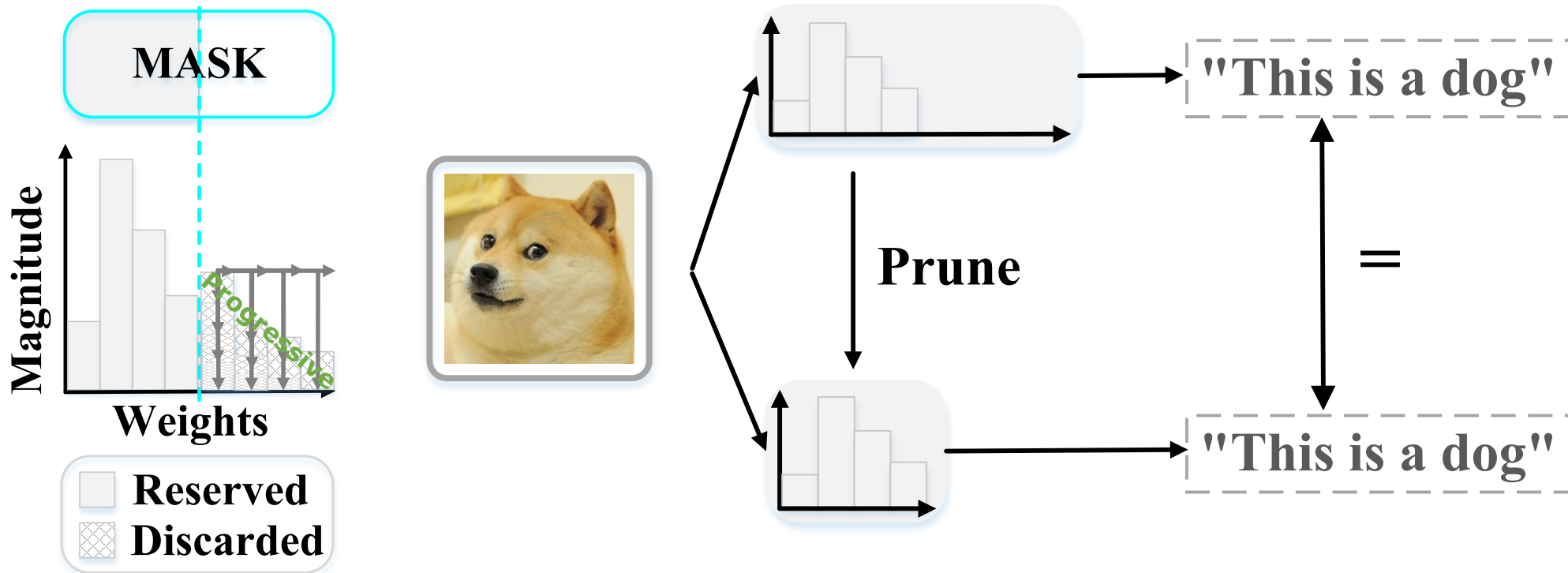# How to prune 50% weights from original models?

🤠 : Don't worry! Iterative pruning will help.

Nice try! How about further improved progressive pruning? : 🤔

How to prune 50% weights from original models?

Progressive pruning eliminates the weight gap!: 😎

🥳 : Cheers to doggy still being doggy after pruning!

Overall Performance

Method	Visual Reason	Image Caption	Visual QA	Retrieval		Image Classification
				COCO	Flickr	
Original Model	83.1	23.8	77.5	81.9	96.8	79.9
Mask-based Pruning	$76.4_{\downarrow 6.7}$	$21.0_{\downarrow 2.8}$	$71.6_{\downarrow 5.9}$	$61.7_{\downarrow 20}$	$78.9_{\downarrow 18}$	$77.9_{\downarrow 2.0}$
UPop (Ours)	$\mathbf{81.1}_{\downarrow 2.0}$	$\mathbf{23.3}_{\downarrow 0.5}$	$\mathbf{76.3}_{\downarrow 1.2}$	$\mathbf{77.4}_{\downarrow 4.5}$	$\mathbf{94.0}_{\downarrow 2.8}$	$\mathbf{78.9}_{\downarrow 1.0}$

At 2× compression for example. All metrics are the higher the better.

Also competitive on uni-modal tasks
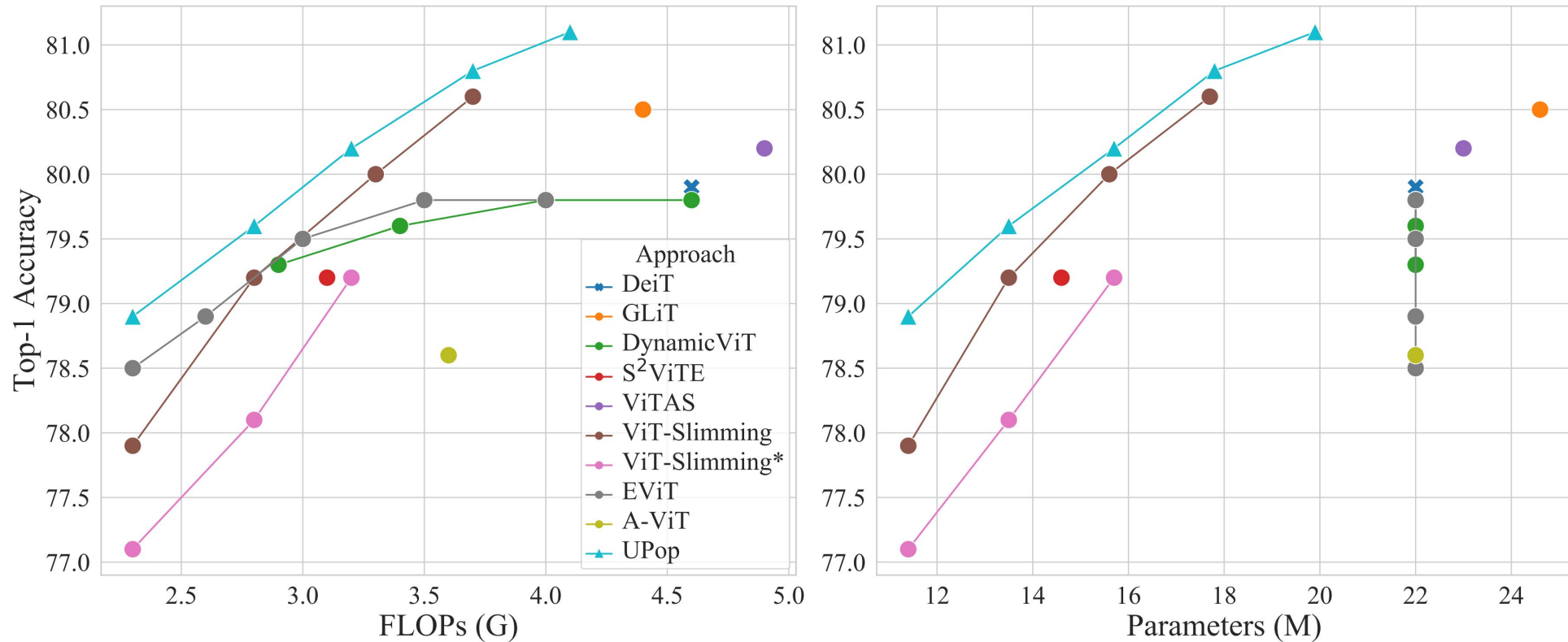
UPop (marked with the blue triangle) achieves better performance on both Accuracy-FLOPs and Accuracy-Parameter trade-offs.

UPop: Unified and Progressive Pruning for Compressing Vision-Language Transformers

Link to paper   Link to code   Link to website

Please refer to above links for more details.