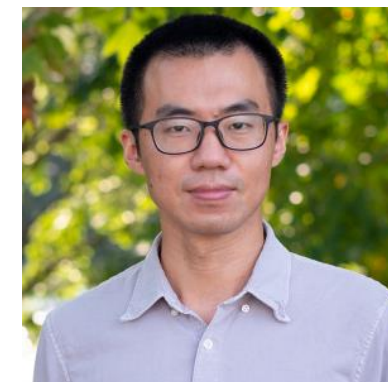
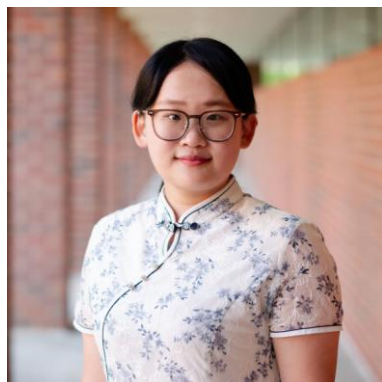
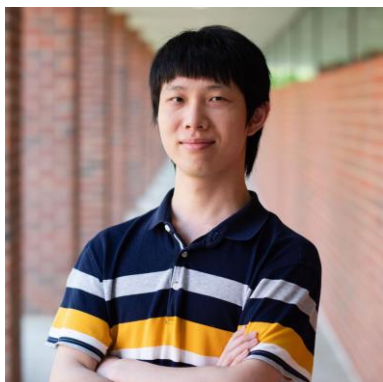




BalancEdit: Dynamically Balancing the Generality- Locality Trade-off in Multi-modal Model Editing

ICML | 2025

Dongliang Guo, Mengxuan Hu, Zihan Guan, Thomas Hartvigsen, Sheng Li



University of Virginia



Is MLLM Always Right?

- **Fact is changing over time**

The training data of the large language model is fixed at a certain point in time, and as time goes by, the internal knowledge of the model will become outdated.



Who owns it?



Is MLLM Always Right?

- **Fact is changing over time**

The training data of the large language model is fixed at a certain point in time, and as time goes by, the internal knowledge of the model will become outdated.



Who owns it?

Jack Dorsey ✗

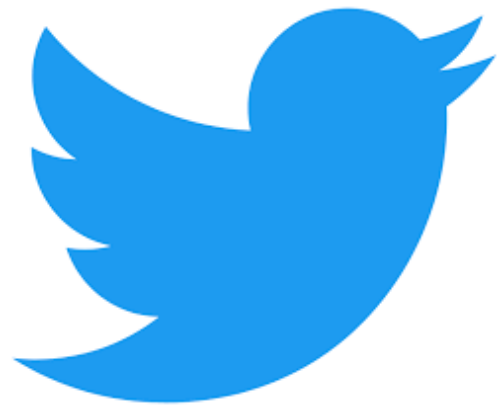




Is MLLM Always Right?

- **Fact is changing over time**

The training data of the large language model is fixed at a certain point in time, and as time goes by, the internal knowledge of the model will become outdated.



Who owns it?

Jack Dorsey ✗



Elon Musk ✓



How to change the model behaviors?

- **Finetune or Re-train**

Example: **Llama**

- Train for **21 days**
- **2048 A100 GPUs**
- **Over \$2.4M**

It is too costly and intractable.



How to change the model behaviors?

- **Model Editing**

Focuses on specific, targeted modifications to a pre-trained model to correct errors, incorporate new knowledge **without retraining the entire model.**

- **Update** the target knowledge (**Reliability**)
- **Influence** related knowledge (**Generality**)
- **Protect** correct knowledge (**Locality**)



How to ensure the influence scope of a fact?

- **Fact has different scope**

Example 1: <Persian cat breed is named as Kitty now>

Example 2: <Cat species is named as Kitty now>

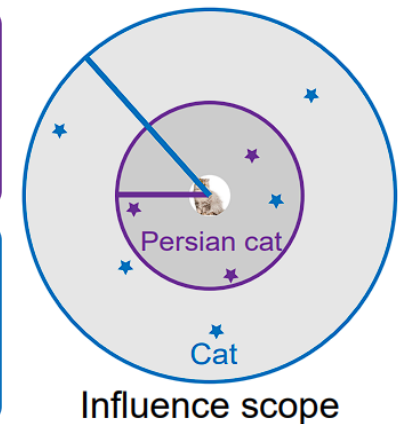


Q: What is the name of the **breeds** in the image?

A: Persian cat \Rightarrow Kitty ✓




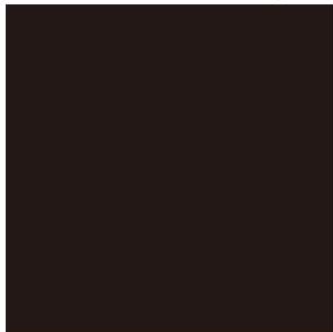
Q: What is the name of the **species** in the image?

A: Cat \Rightarrow Kitty ✓





Can existing editing method differentiate it?

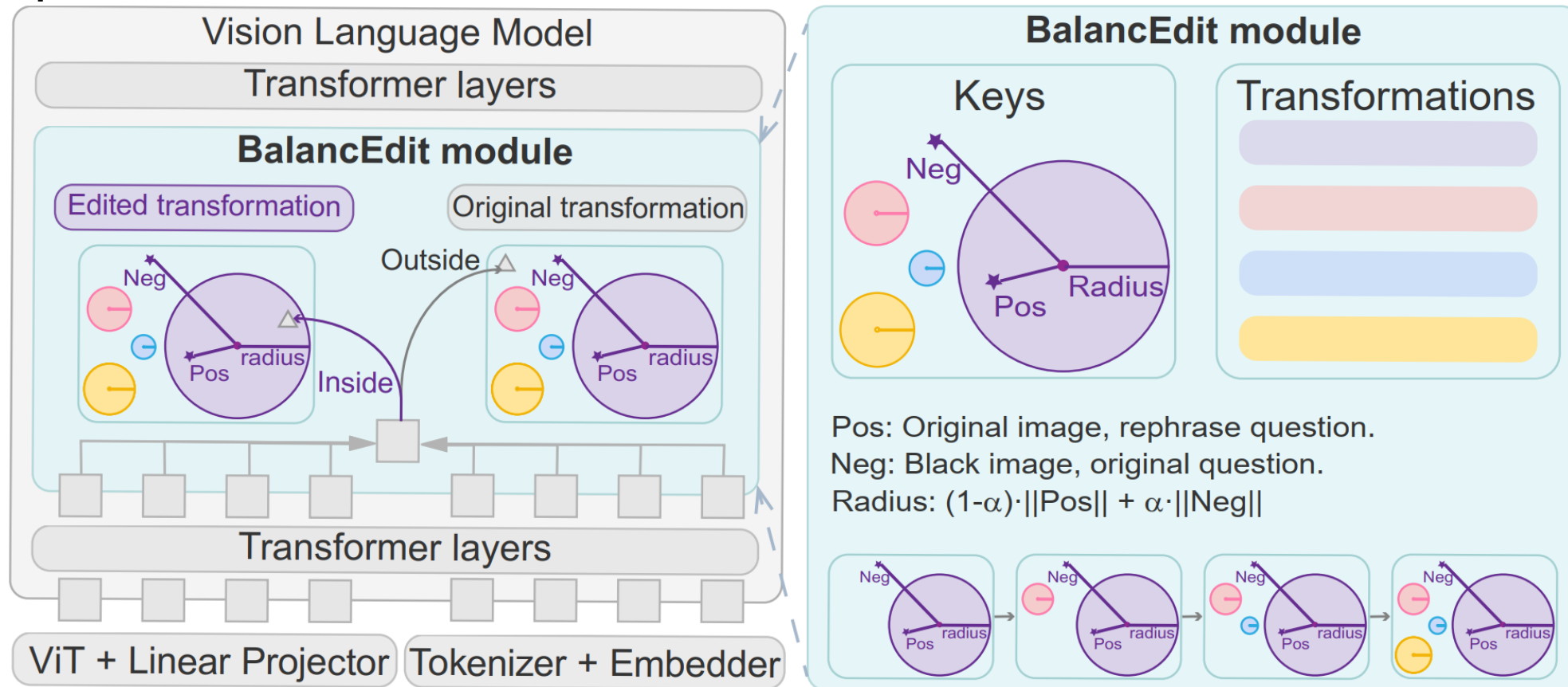
	Original Image	Related Image	Unrelated Image	Black Image
				
Question	What brand is this computer?			
Target	hp → lenovo			
Base	hp	hp	dell	black
IKE	lenovo	lenovo	lenovo	lenovo
MEND	lenovo	lenovo	lenovo	lenovo
GRACE	lenovo	hp	dell	black
Ours	lenovo	lenovo	dell	black

They may fail even if a black image is given for the question.



How to dynamically determine the scope?

- **BalancEdit**: Dynamically determine the Generality-Locality equilibrium





How to dynamically determine the scope?

Example: <Cat species is named as Kitty now>

What is its
species?



Center





How to dynamically determine the scope?

Step 1: Generate positive and negative sample

Altered text question

What is its
species?

What *category*
is it?

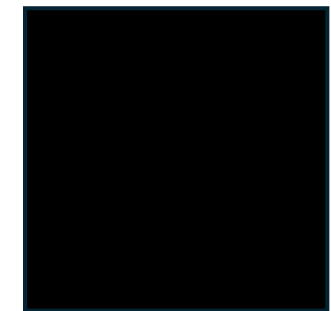


Center

Positive sample

Altered image input

What is its
species?



Negative sample





How to dynamically determine the scope?

Step 2: Determine the radius and learn new knowledge

What is its
species?



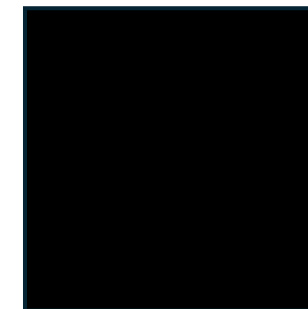
Center

What **category**
is it?



Positive sample

What is its
species?



Negative sample



$$\text{Radius} = a * \text{positive} + (1-a) * \text{negative}$$



How to dynamically determine the scope?

Step 3: Evaluation

What is its
species?



Center

What **category**
is it?



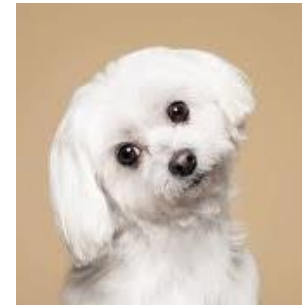
Positive sample

What is its
species?



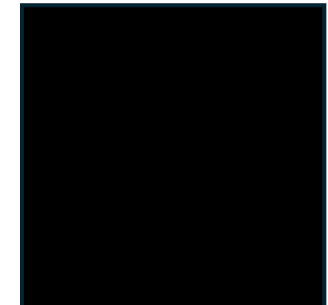
Generality sample

What is its
species?



Locality sample

What is its
species?



Negative sample

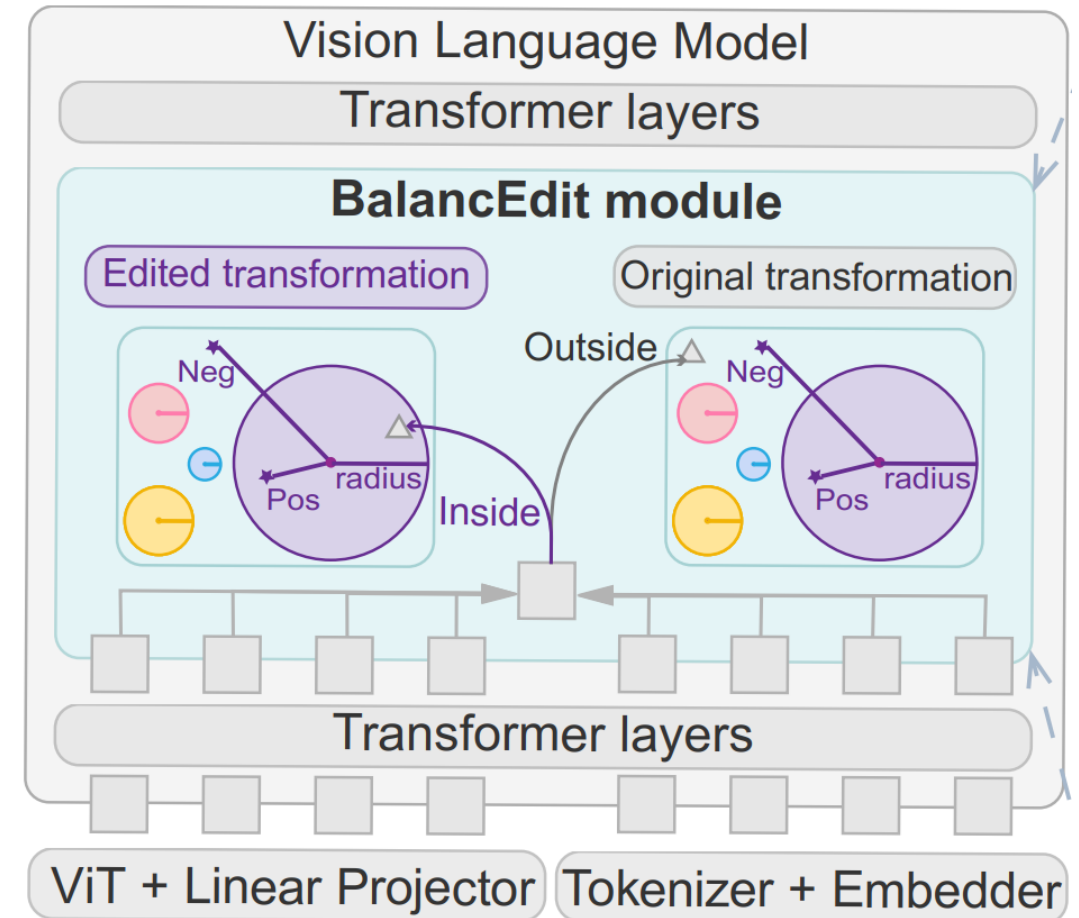


$$\text{Radius} = a * \text{positive} + (1-a) * \text{negative}$$



How to inference the MLLM after editing?

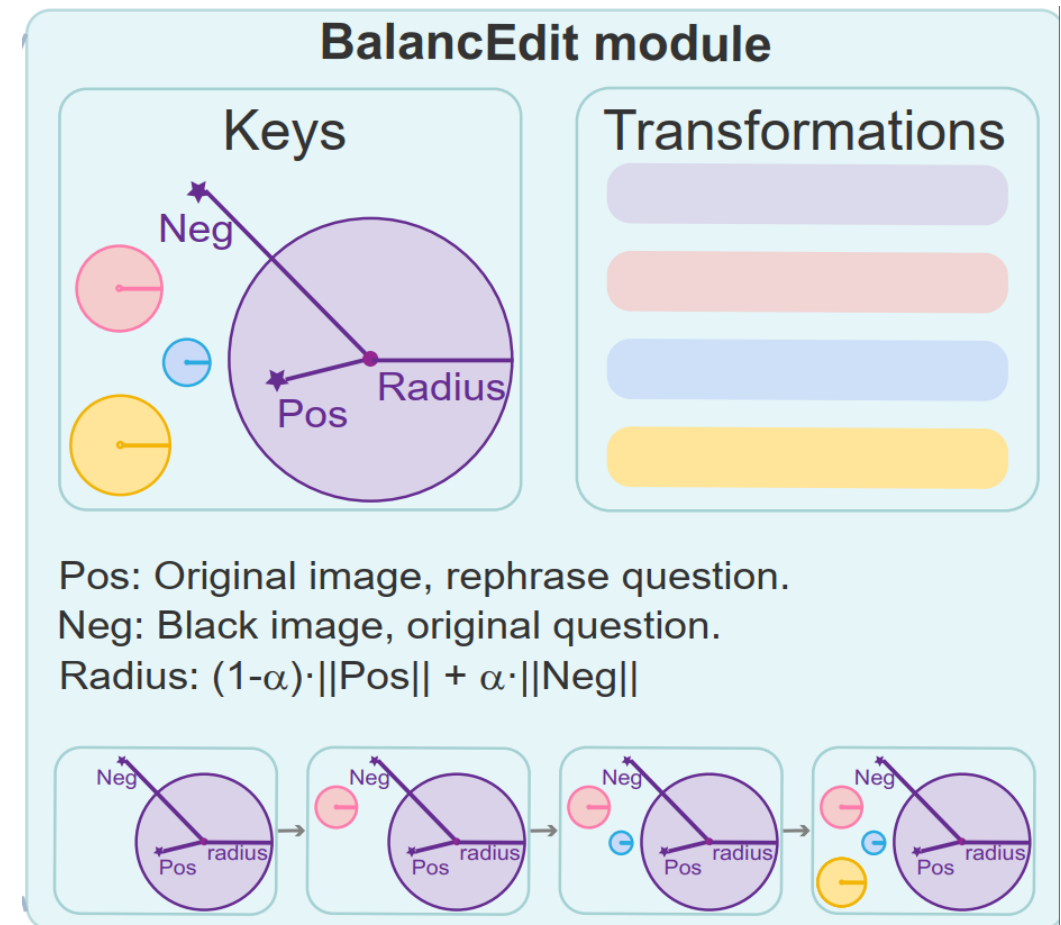
- The input is **related** to the new knowledge.
 - Find the target new knowledge
 - Invoke the edited transformation layer
- The input is **not related** to the new knowledge.
 - Go through the original layer without any changes





How to continually edit MLLM?

- Adding memory to the codebook
- Dynamically adjusting the radius based on knowledge overlapping.





How is the Performance?

- Outperforms** existing methods on Generality-Locality trade-off

Dataset	Method	Pretrain	Backbone									
			miniGPT4					BLIP-2 OPT				
			Acc↑	T-Gen↑	I-Gen↑	Loc↑	HM↑	Acc↑	T-Gen↑	I-Gen↑	Loc↑	HM↑
MMEDIT	Base	✗	15.04	14.21	13.56	NA	NA	8.50	8.52	6.89	NA	NA
	FT	✗	96.53	95.88	96.20	3.20	9.00	99.96	99.41	97.05	0.27	0.80
	IKE	✓	100.00	95.57	100.00	15.47	20.07	99.83	94.47	99.58	11.96	28.77
	MEND	✓	98.39	96.58	97.77	68.82	85.43	97.23	95.86	96.81	69.40	85.29
	GRACE	✗	79.82	74.49	70.11	91.66	77.72	74.27	62.90	35.24	90.26	54.19
	BalancEdit (Ours)	✗	100.00	99.90	98.91	71.74	88.08	100.00	99.16	90.30	80.04	89.14
OKEDIT	Base	✗	30.42	45.40	72.21	NA	NA	14.35	13.96	15.22	NA	NA
	FT	✗	99.69	99.45	99.38	5.52	14.90	99.97	99.54	96.77	0.43	1.27
	IKE	✓	99.71	97.78	99.76	17.45	38.68	99.35	94.20	99.66	13.29	31.28
	MEND	✓	94.44	90.80	95.39	36.20	61.07	90.82	82.82	88.25	28.89	51.70
	GRACE	✗	87.84	28.31	29.46	99.99	37.84	54.13	50.67	28.30	94.48	45.69
	BalancEdit (Ours)	✗	100.00	99.87	76.46	53.14	71.58	100.00	98.89	65.38	61.18	71.85

High Editing Accuracy

Minimal trade-off



How is the Performance?

- **Outperforms** existing methods on **Generality-Locality trade-off**
- **Outperforms** existing methods on **Sequential Edits**

	Sequential	Acc↑	T-Gen↑	I-Gen↑	Loc↑	HM↑
FT	✗	99.25	99.21	98.64	0.74	2.18
IKE	✗	100.00	96.86	100.00	16.91	37.75
MEND	✗	93.74	89.98	95.38	37.49	62.14
GRACE	✗	87.78	25.96	24.21	99.99	33.39
BalancEdit (Ours)	✗	100.00	100.00	72.31	54.40	71.07
BalancEdit (Ours)	✓	100.00	99.70	72.29	46.25	65.95

Robust on Sequential edits



How is the Performance?

- **Outperforms** existing methods on **Generality-Locality trade-off**
- **Outperforms** existing methods on **Sequential Edits**
- **Time and Data Efficient**

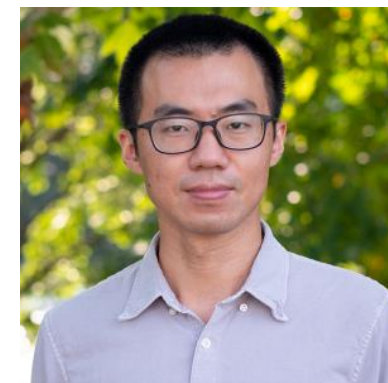
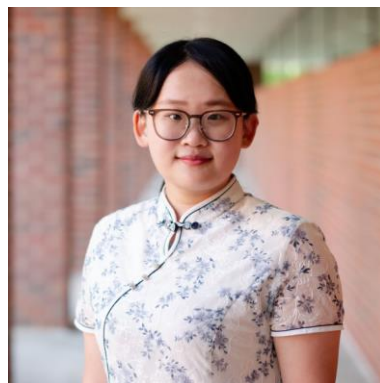
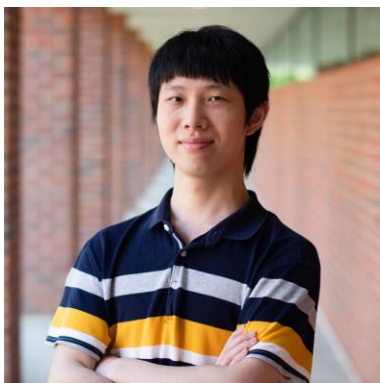
	Training time (h)	Editing time (s)
FT	0	3.91
IKE	12	0.38
MEND	22	1.48
GRACE	0	32.67
BalancEdit	0	8.04

No pre-training needed *Efficient edits*



BalancEdit: Dynamically Balancing the Generality- Locality Trade-off in Multi-modal Model Editing

Dongliang Guo, Mengxuan Hu, Zihan Guan, Thomas Hartvigsen, Sheng Li



Open to Work

Get in touch!

Dongliang.guo@virginia.edu

**Check the paper
and Code**

