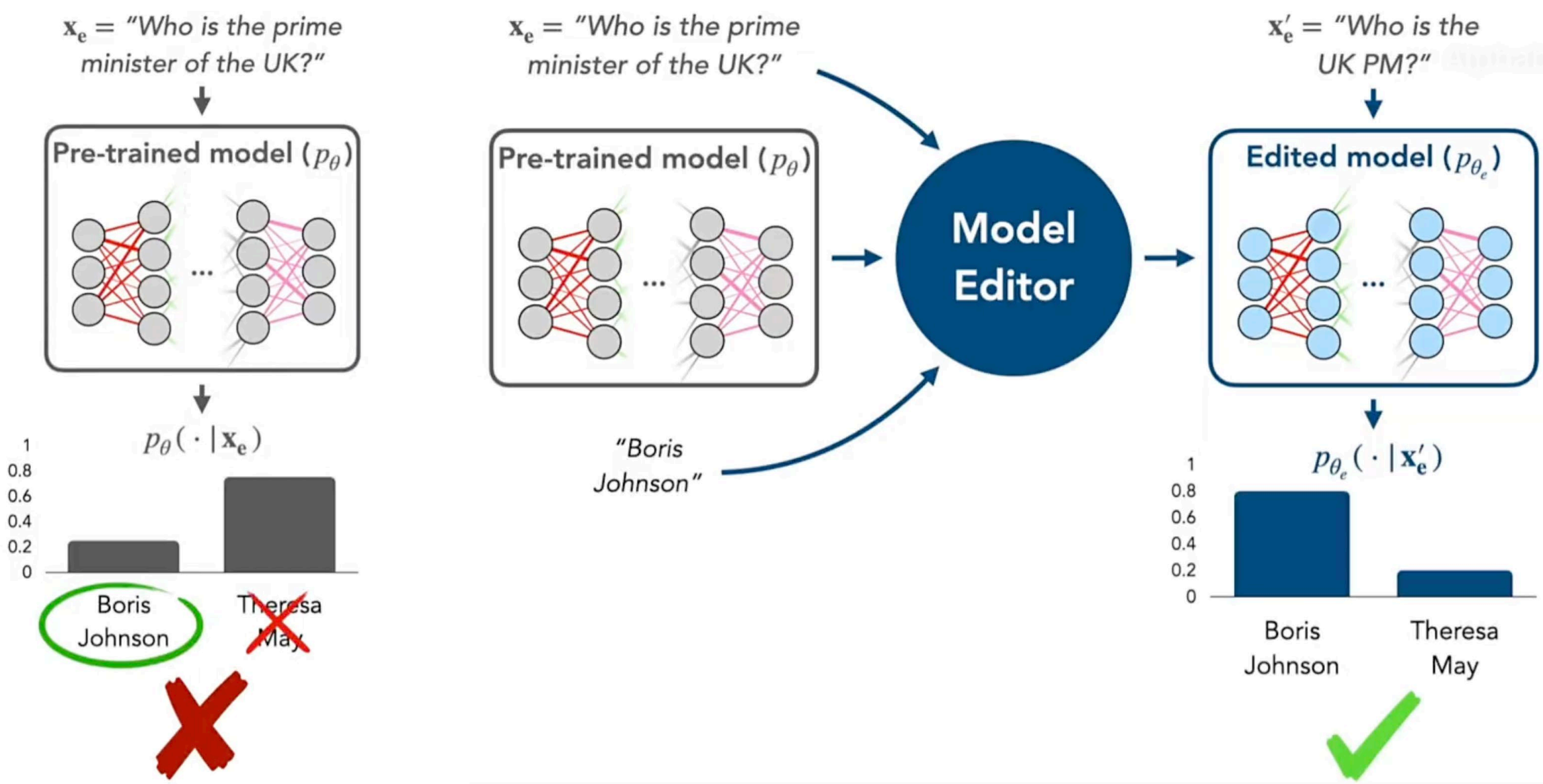
Can We Edit Multimodal Large Language Models?

Siyuan Cheng[♣♥♡*], **Bozhong Tian**[♣♥*], **Qingbin Liu**[♡], **Xi Chen**[♡†],
Yongheng Wang[♢], **Huajun Chen**[♣♥♠], **Ningyu Zhang**[♣♥†]
[♣] Zhejiang University [♥] Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph
[♠] Donghai Laboratory [♡] Platform and Content Group, Tencent [♢] Zhejiang Laboratory
{sycheng,tbozhong,huajunsir,zhangningyu}@zju.edu.cn
{qingbinliu,jasonxchen}@tencent.com,wangyh@zhejianglab.com

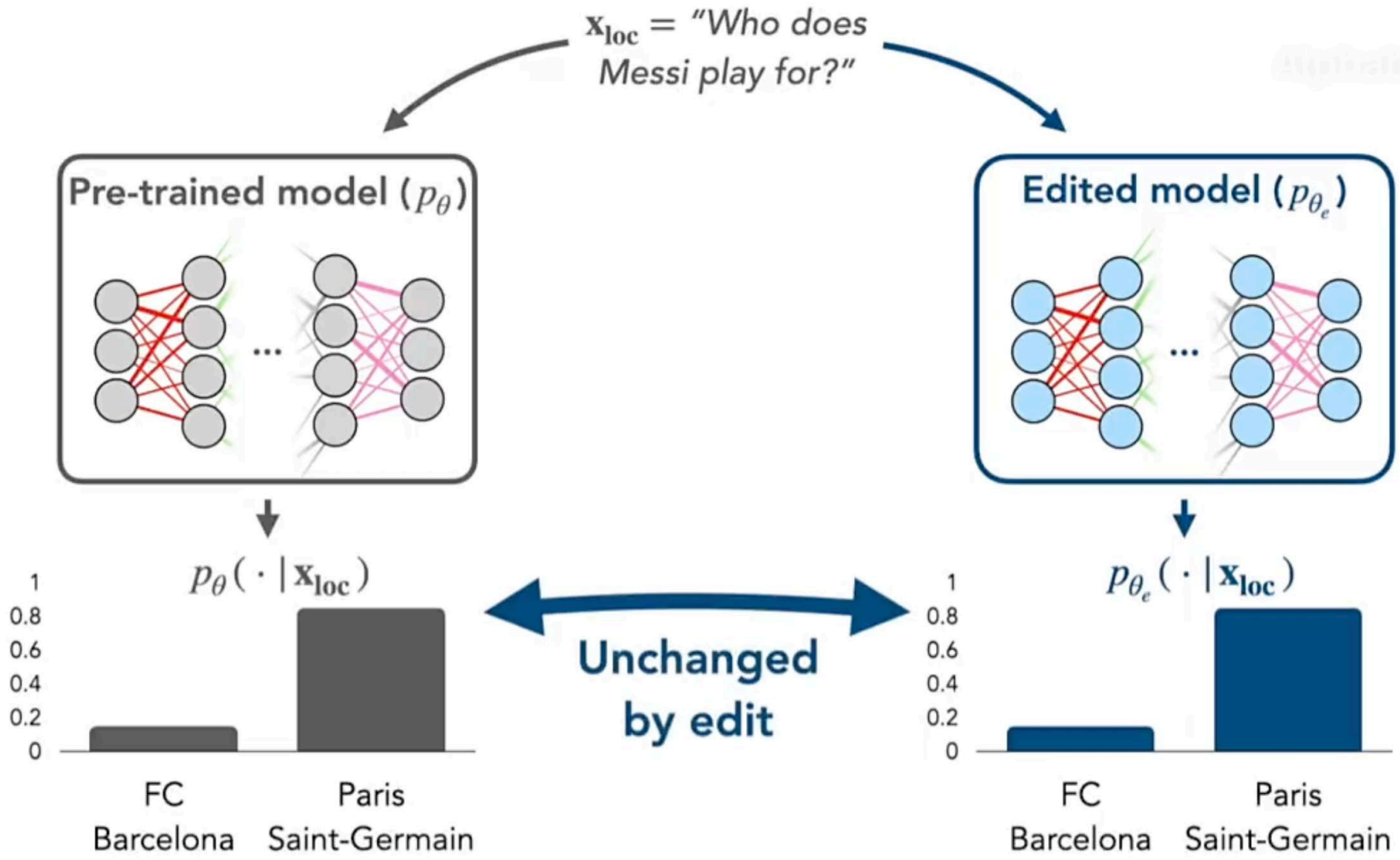
1.Constructing a new benchmark, dubbed **MMEdit**, for editing and evaluation.

2.Conducting experiments involving various model editing baselines.

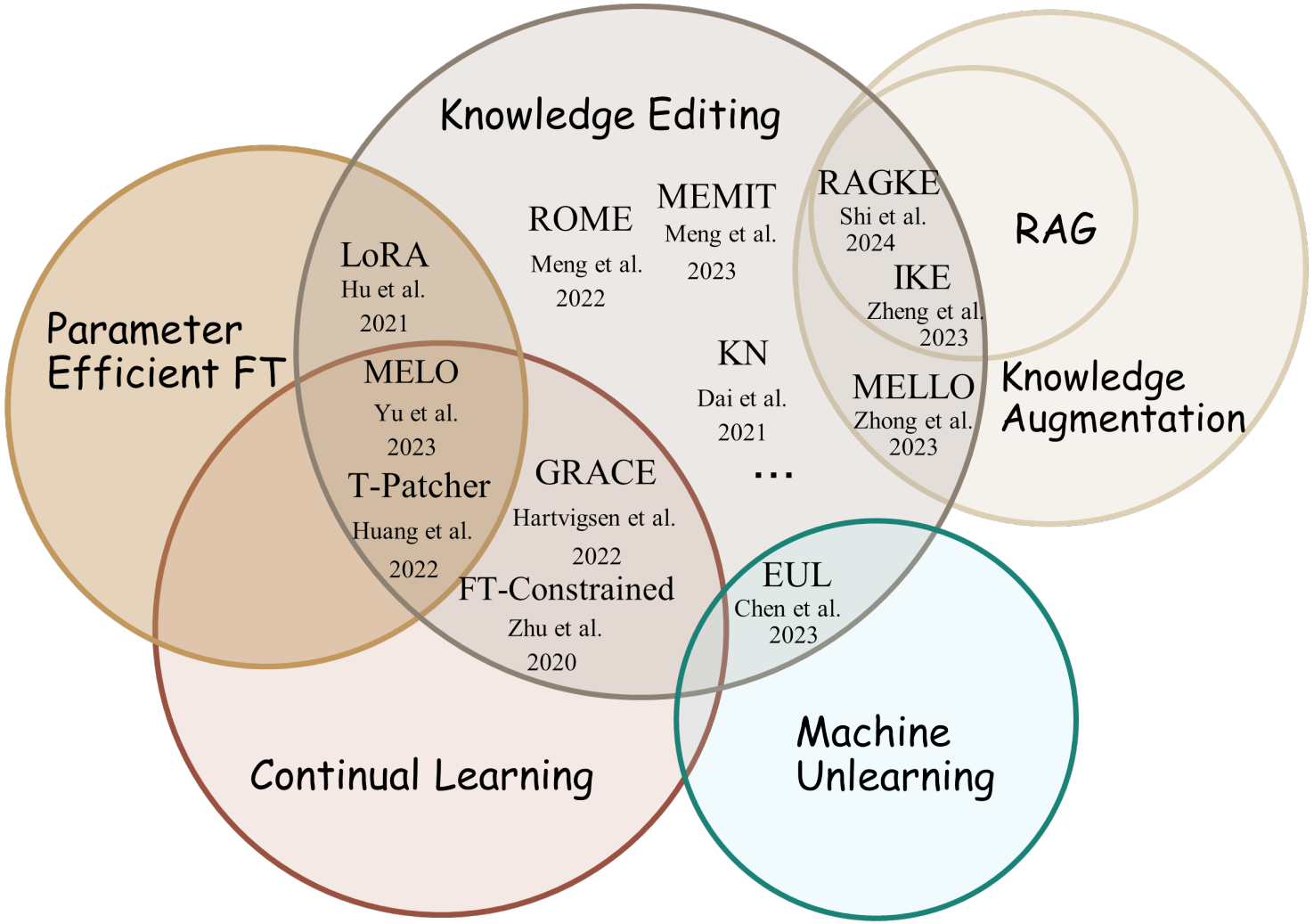
Overview


$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
Pre-trained model (p_θ)
$p_\theta(\cdot|\mathbf{x_e})$
1
0.8
0.6
0.4
0.2
0
Boris Johnson
Theresa May
$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
Pre-trained model (p_θ)
"Boris Johnson"
Model Editor
$\mathbf{x'_e}$ = "Who is the UK PM?"
Edited model (p_{θ_e})
$p_{\theta_e}(\cdot|\mathbf{x'_e})$
1
0.8
0.6
0.4
0.2
0
Boris Johnson
Theresa May

Overview

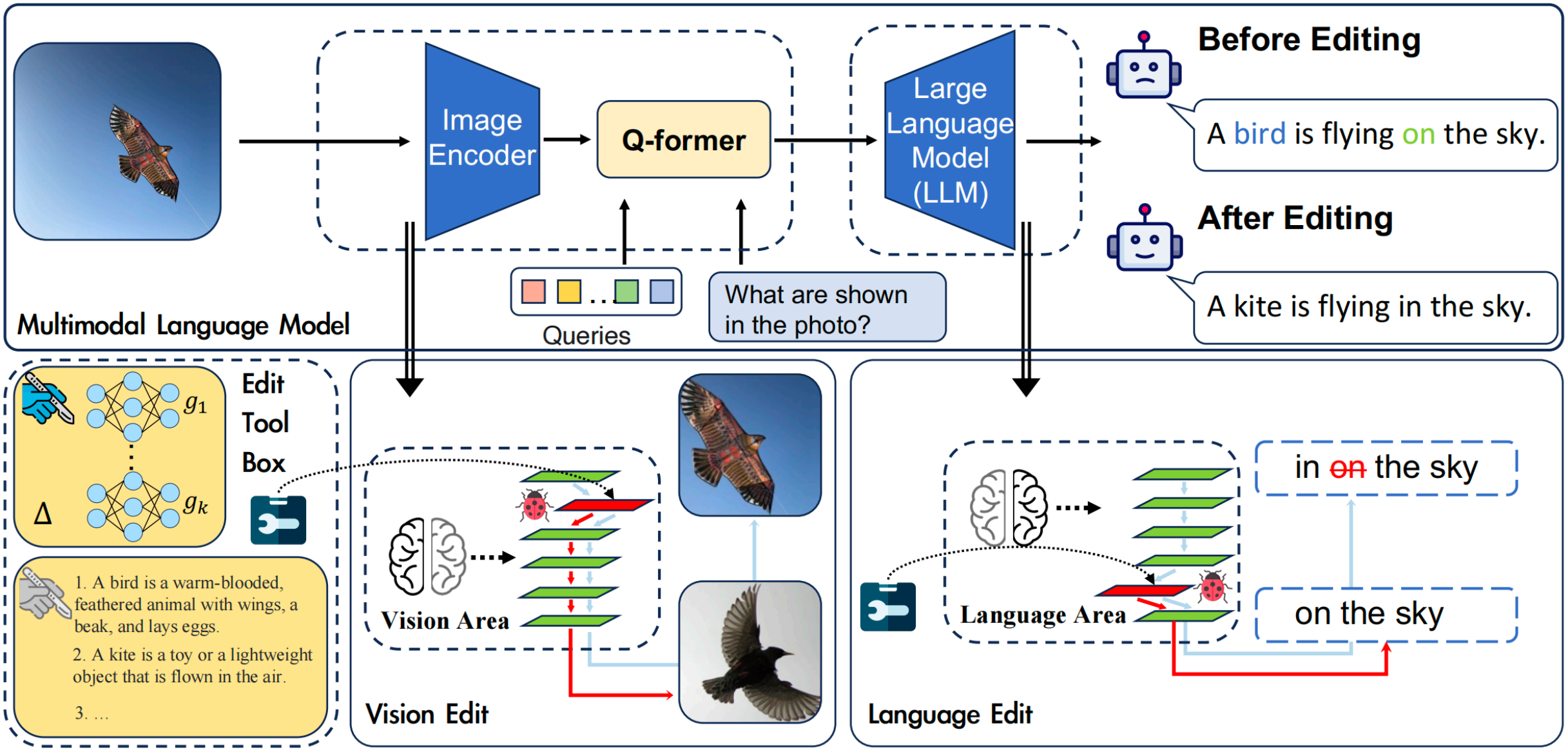

$\mathbf{x}_{\text{loc}}$ = "Who does Messi play for?"
Pre-trained model (p_θ)
Edited model (p_{θ_e})
$p_\theta(\cdot \mid \mathbf{x}_{\text{loc}})$
$p_{\theta_e}(\cdot \mid \mathbf{x}_{\text{loc}})$
1
0.8
0.6
0.4
0.2
0
FC Barcelona
Paris Saint-Germain
Unchanged by edit

Overview

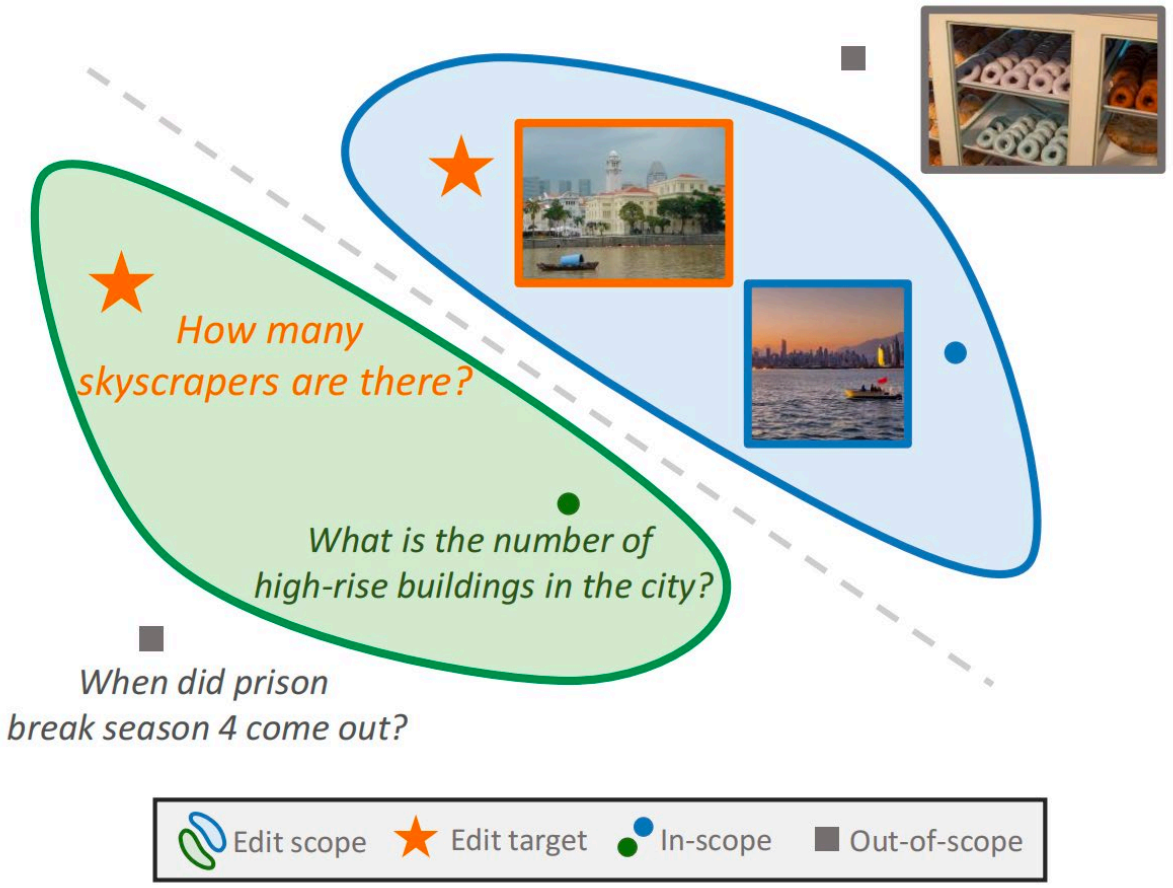


Comparison of different technologies

Editing Multimodal LLMs



Editing Multimodal LLMs



Reliability

$$\mathcal{M}_{rel} = \mathbb{E}_{(i_e,x_e,y_e)\sim\mathcal{D}_{\text{edit}}} \left[\mathbb{1}_{f(i_e,x_e;\theta_e(i_e,x_e,y_e))=y_e} \right]$$

Locality

$$\mathcal{M}_{loc}^{Text} = \mathbb{E}_{\substack{(i_e,x_e,y_e)\sim\mathcal{D}_{\text{edit}} \\ (x,y)\sim\mathcal{D}_{\text{loc-t}}}} \left[\mathbb{1}_{f(x;\theta_e(i_e,x_e,y_e))=f(x,\theta)} \right]$$

$$\mathcal{M}_{loc}^{Img} = \mathbb{E}_{(i_v,x_v,y_v)\sim\mathcal{D}_{\text{loc-v}}} \left[\mathbb{1}_{f(i_v,x_v;\theta_e)=f(i_v,x_v;\theta)} \right]$$

Generality

$$\mathcal{M}_{gen}^{Text} = \mathbb{E}_{(x_r)\sim\mathcal{N}(x_e)} \left[\mathbb{1}_{f(i_e,x_r;\theta_e)=f(i_e,x_e;\theta_e)} \right]$$

(

$$\mathcal{M}_{gen}^{Img} = \mathbb{E}_{(i_r)\sim\mathcal{N}(i_e)} \left[\mathbb{1}_{f(i_r,x_e;\theta_e)=f(i_e,x_e;\theta_e)} \right]$$

MMEdit

1.Reliability Dataset: VQAv2; COCO Caption.

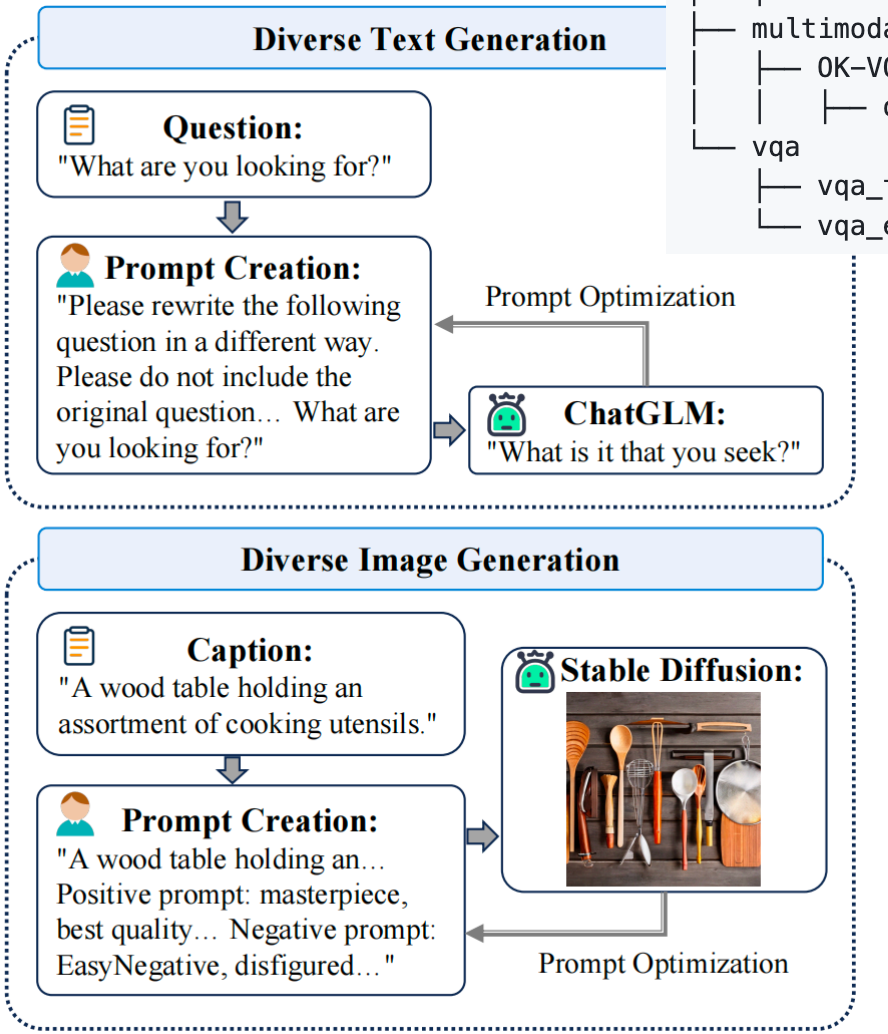
2.Locality Dataset:

(1)Textual Locality Dataset: NQ (previously used in MEND)

(2)MultiModal Locality Dataset: OK-VQA

3.Generality Dataset:

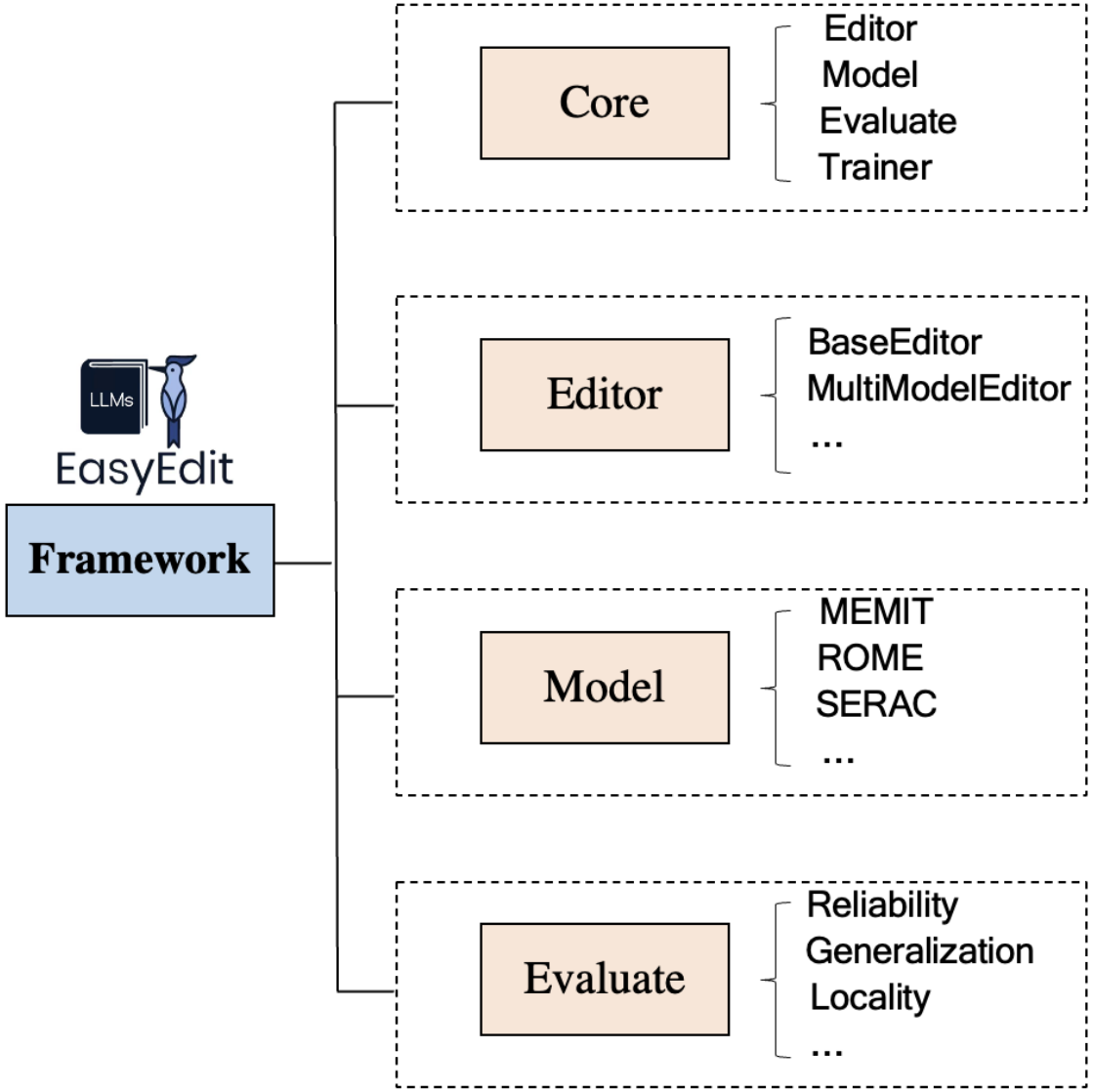
```
editing-data
├── caption
│   ├── caption_train_edit.json
│   └── caption_eval_edit.json
├── locality
│   ├── NQ dataset
│   │   ├── train.json
│   │   └── validation.json
├── multimodal_locality
│   ├── OK-VQA dataset
│   │   ├── okvqa_loc.json
└── vqa
    ├── vqa_train.json
    └── vqa_eval.json
```

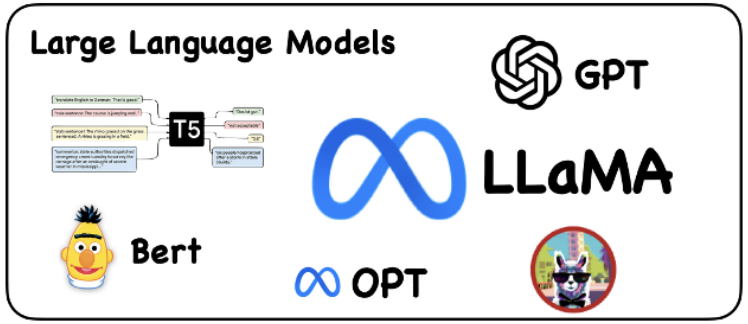


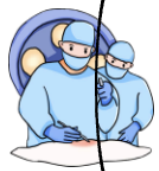
The statistic of datasets for the E-VQA and E-IC sub-tasks.

TASK	Train	Test	L-Locality	M-Locality
E-VQA	6,346	2,093	4,289	5,046
E-IC	2,849	1,000	4,289	5,046

EasyEdit

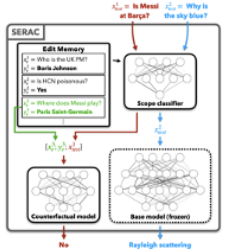




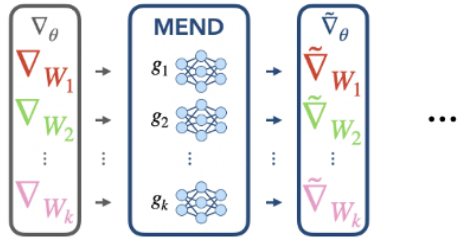


Model Editing

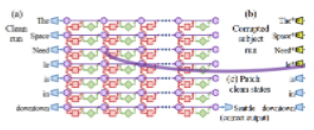
various methods



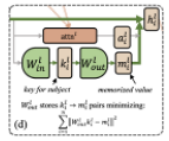
SERAC(ICML22)



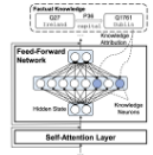
MEND(ICLR22)



ROME(NeurIPS22)

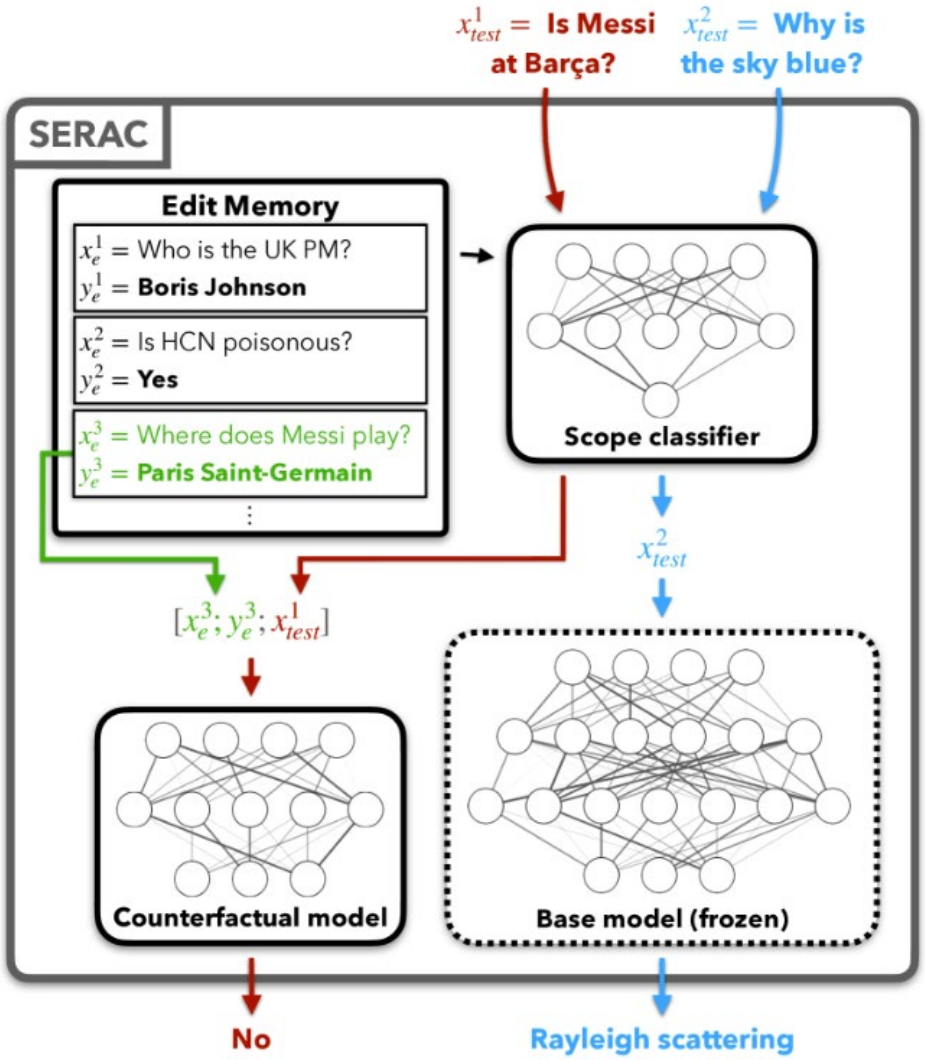


MEMIT(ICLR23)



Knowledge Neuron(ACL22)

SERAC



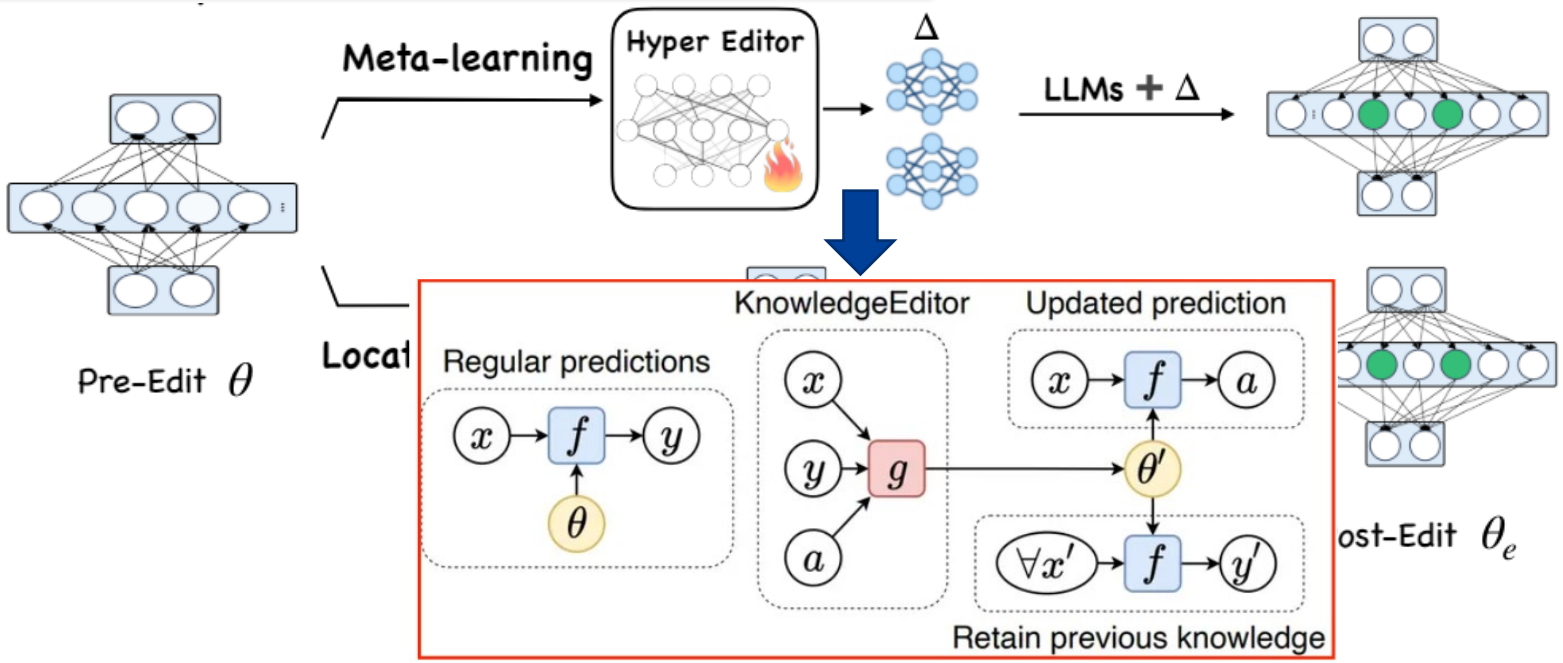

```yaml
inner_params:
- opt_model.model.decoder.layers.29.fc1.weight
- opt_model.model.decoder.layers.29.fc2.weight
- opt_model.model.decoder.layers.30.fc1.weight
- opt_model.model.decoder.layers.30.fc2.weight
- opt_model.model.decoder.layers.31.fc1.weight
- opt_model.model.decoder.layers.31.fc2.weight

# Method
alg: SERAC_MULTI
alg_name: SERAC_MULTI
```

		Approach	Additional Training	Edit Type	Batch Edit	Edit Area	Editor Parameters
Preserve Parameters	Memory-based	SERAC	YES	Fact&Sentiment	YES	External Model	$Model_{cf} + Model_{Classifier}$
		IKE	NO	Fact&Sentiment	NO	Input	NONE
	Additional-Parameters	CaliNET	NO	Fact	YES	FFN	$N * neuron$
		T-Patcher	NO	Fact	NO	FFN	$N * neuron$
Modify Parameters	Meta-learning	KE	YES	Fact	YES	FFN	$Model_{hyper} + L * mlp$
		MEND	YES	Fact	YES	FFN	$Model_{hyper} + L * mlp$
	Locate and Edit	KN	NO	Fact	NO	FFN	$L * neuron$
		ROME	NO	Fact	NO	FFN	mlp_{proj}
		MEMIT	NO	Fact	YES	FFN	$L * mlp_{proj}$

MEND

MEND

```yaml
alg_name: "MEND"
name: hugging_cache/opt-2.7b
model_name: blip2
model_class: Blip2OPT
tokenizer_class: GPT2Tokenizer
tokenizer_name: hugging_cache/opt-2.7b
inner_params:
- opt_model.model.decoder.layers.29.fc1.weight
- opt_model.model.decoder.layers.29.fc2.weight
- opt_model.model.decoder.layers.30.fc1.weight
- opt_model.model.decoder.layers.30.fc2.weight
- opt_model.model.decoder.layers.31.fc1.weight
- opt_model.model.decoder.layers.31.fc2.weight
```



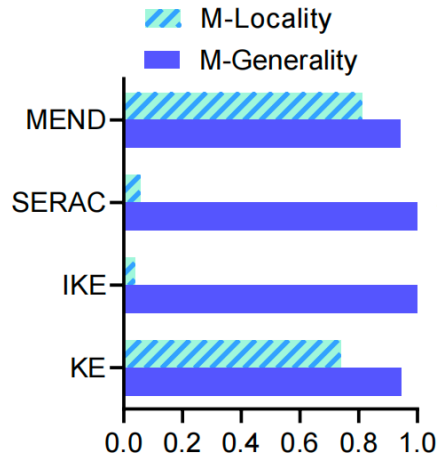
KE

```yaml
inner_params:
- transformer.h.25.mlp.fc_in.weight
- transformer.h.25.mlp.fc_out.weight
- transformer.h.26.mlp.fc_in.weight
- transformer.h.26.mlp.fc_out.weight
- transformer.h.27.mlp.fc_in.weight
- transformer.h.27.mlp.fc_out.weight


# Method
alg: KE
```

		Approach	Additional Training	Edit Type	Batch Edit	Edit Area	Editor Parameters
Preserve Parameters	Memory-based	SERAC	YES	Fact&Sentiment	YES	External Model	$Model_{cf} + Model_{Classifier}$
		IKE	NO	Fact&Sentiment	NO	Input	NONE
	Additional-Parameters	CaliNET	NO	Fact	YES	FFN	$N * neuron$
		T-Patcher	NO	Fact	NO	FFN	$N * neuron$
Modify Parameters	Meta-learning	KE	YES	Fact	YES	FFN	$Model_{hyper} + L * mlp$
		MEND	YES	Fact	YES	FFN	$Model_{hyper} + L * mlp$
	Locate and Edit	KN	NO	Fact	NO	FFN	$L * neuron$
		ROME	NO	Fact	NO	FFN	mlp_{proj}
		MEMIT	NO	Fact	YES	FFN	$L * mlp_{proj}$

Experiments

	Method	Editing VQA				Editing Image Caption			
		Reliability ↑	T-Generality ↑	T-Locality ↑	M-Locality ↑	Reliability ↑	T-Generality ↑	T-Locality ↑	M-Locality ↑
	BLIP-2 OPT								Size: 3.8B
Base Methods	Base Model	0.00/0.00	0.00/0.00	100.0	100.0	0.00/13.33	0.00/13.33	100.0	100.0
	FT (vision block)	60.56/60.98	49.79/50.22	100.0	8.47	18.94/79.30	5.86/72.06	100.0	8.40
	FT (last layer)	57.66/57.97	46.70/49.22	21.67	3.06	16.60/61.08	3.50/52.14	24.96	7.12
Model Editing	Knowledge Editor	85.28/92.64	84.23/92.22	90.31	52.48	0.30/50.52	0,10/48.96	88.31	49.52
	In-Context Editing	99.71 / 99.95	91.59/93.93	48.79	2.53	93.80 / 96.70	69.40/78.20	48.95	2.95
	SERAC	99.90 / 99.97	99.90 / 99.99	100.0	2.91	98.90 / 99.92	98.90 / 99.91	99.98	7.52
	MEND	98.51/99.40	97.51 / 98.80	99.94	96.65	80.00/96.11	78.10 / 95.82	94.54	70.84
	MiniGPT-4								Size: 7.3B
Base Methods	Base Model	0.00/0.00	0.00/0.00	100.0	100.0	0.00/31.25	0.00/37.50	100.0	100.0
	FT (vision block)	36.3/72.52	0.3/59.99	100.0	9.29	3.10/60.77	0.00/56.12	100.0	8.56
	FT (last layer)	0.10/70.08	0.00/55.69	72.60	15.75	0.00/65.41	0.00/60.09	53.50	12.68
Model Editing	Knowledge Editor	95.37/98.77	92.64/ 97.28	97.31	73.76	75.50/ 97.58	67.80/ 96.59	97.15	69.92
	In-Context Editing	100.0 / 100.0	94.40 /94.89	50.30	3.67	77.00/90.90	51.80/81.60	52.18	4.68
	SERAC	99.90 / 99.96	92.60/95.06	99.90	5.52	97.30 / 99.79	74.60 / 97.73	99.89	7.20
	MEND	96.20/98.80	95.40 / 98.60	98.23	81.08	80.80 /96.55	78.60 /96.08	98.41	75.25



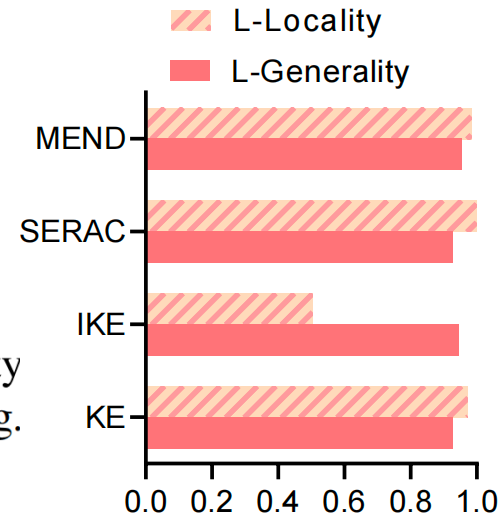


Table 2: Main results on the **MMEdit**. **T-Locality, M-Locality** refer to the textual and multimodal stability **T-Generality** represents textual generality. **Reliability** denotes the (**Exact Match/Accuracy**) of successful editing.

Experiments

MEND

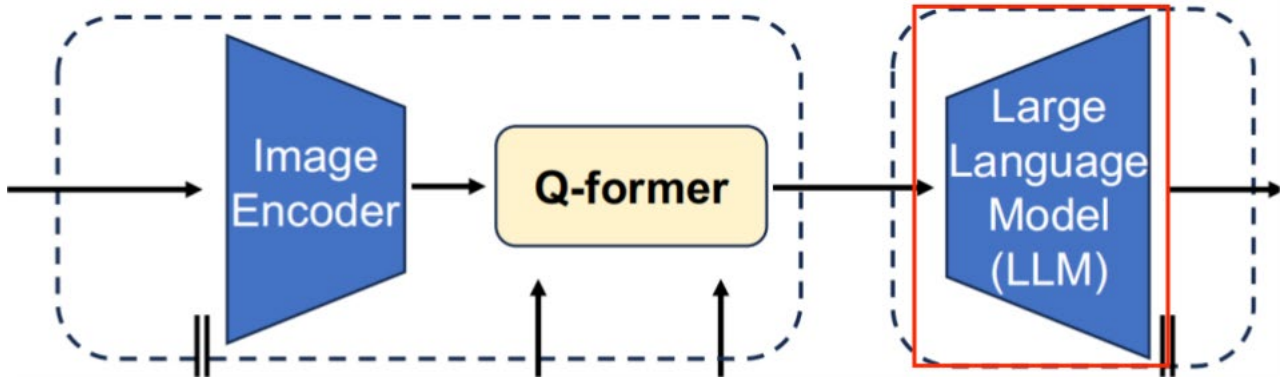
```yaml
alg_name: "MEND"
name: hugging_cache/opt-2.7b
model_name: blip2
model_class: Blip2OPT
tokenizer_class: GPT2Tokenizer
tokenizer_name: hugging_cache/opt-2.7b
inner_params:
- opt_model.model.decoder.layers.29.fc1.weight
- opt_model.model.decoder.layers.29.fc2.weight
- opt_model.model.decoder.layers.30.fc1.weight
- opt_model.model.decoder.layers.30.fc2.weight
- opt_model.model.decoder.layers.31.fc1.weight
- opt_model.model.decoder.layers.31.fc2.weight
```

SERAC

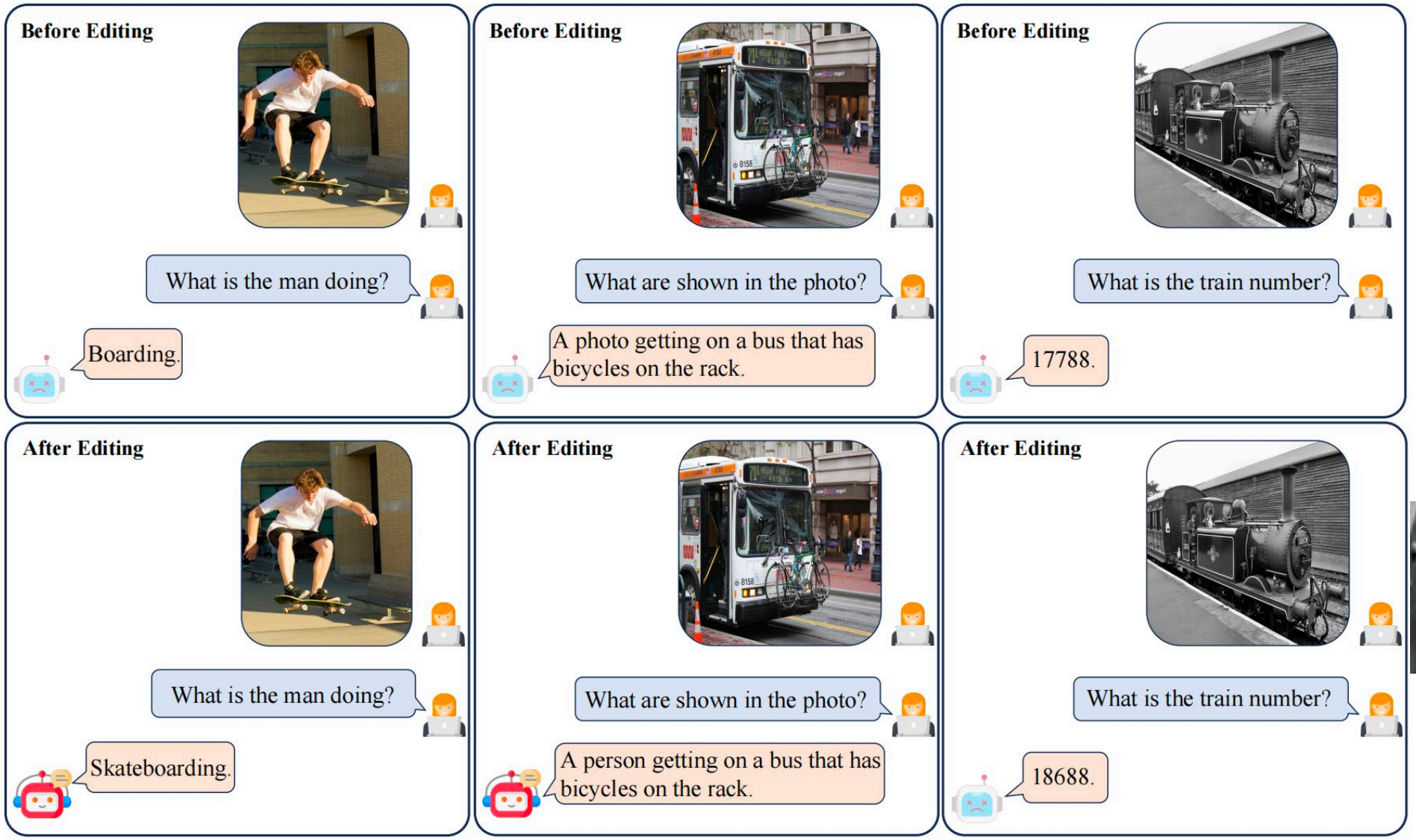
```yaml
inner_params:
- opt_model.model.decoder.layers.29.fc1.weight
- opt_model.model.decoder.layers.29.fc2.weight
- opt_model.model.decoder.layers.30.fc1.weight
- opt_model.model.decoder.layers.30.fc2.weight
- opt_model.model.decoder.layers.31.fc1.weight
- opt_model.model.decoder.layers.31.fc2.weight

# Method
alg: SERAC_MULTI
alg_name: SERAC_MULTI
```

KE

```yaml
inner_params:
- transformer.h.25.mlp.fc_in.weight
- transformer.h.25.mlp.fc_out.weight
- transformer.h.26.mlp.fc_in.weight
- transformer.h.26.mlp.fc_out.weight
- transformer.h.27.mlp.fc_in.weight
- transformer.h.27.mlp.fc_out.weight


# Method
alg: KE
```



Experiments



Experiments

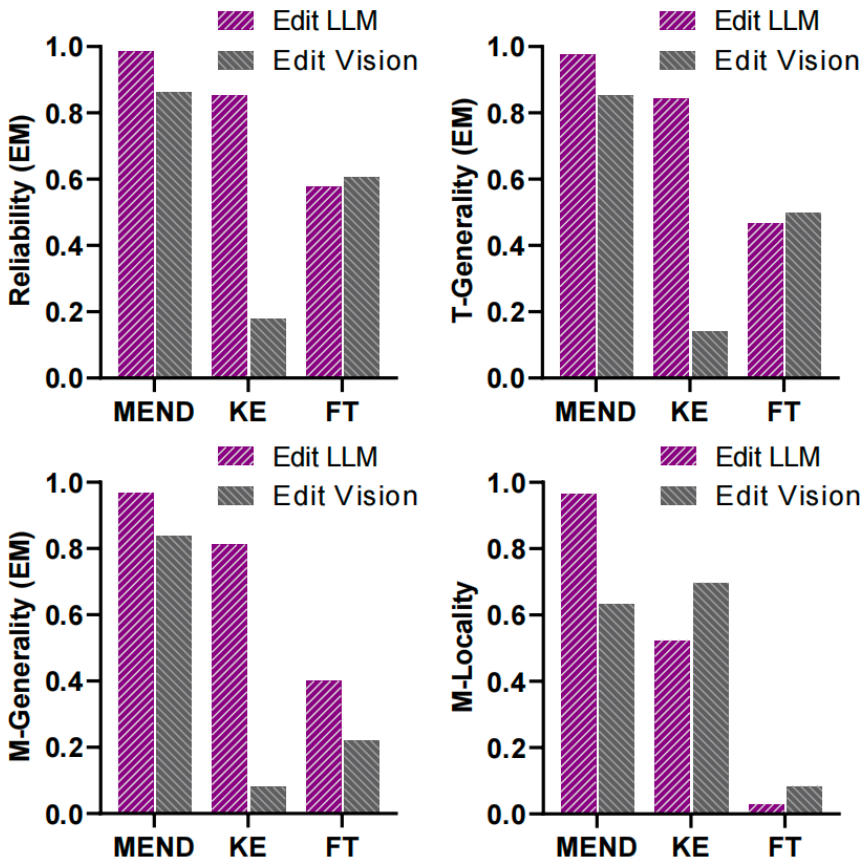
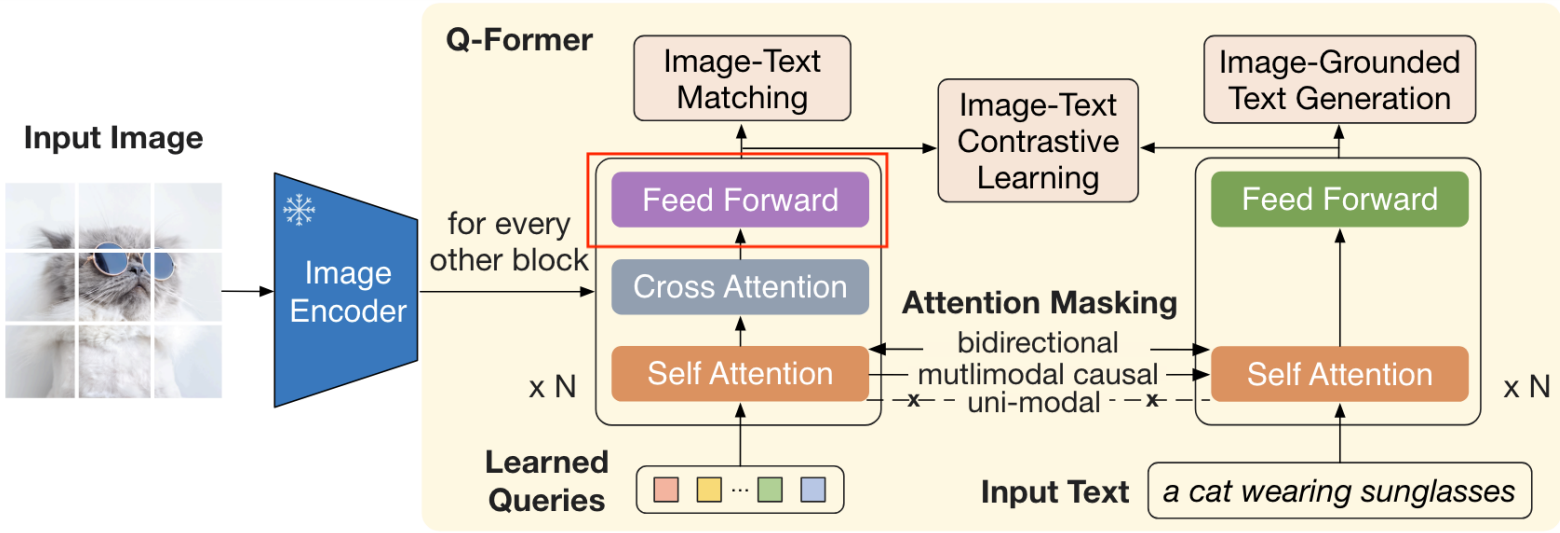


Figure 7: Results of editing different components.




```
4     name: hugging_cache/opt-2.7b
5     model_name: blip2
6     model_class: Blip2OPT
7     tokenizer_class: GPT2Tokenizer
8     tokenizer_name: hugging_cache/opt-2.7b
9     inner_params:
10    - Qformer.bert.encoder.layer.10.output_query.dense.weight
11    - Qformer.bert.encoder.layer.11.output_query.dense.weight
```