



# MMKE-Bench:

## A Multimodal Editing Benchmark for Diverse Visual Knowledge

Yuntao Du<sup>\* 1</sup> Kailin Jiang<sup>\* 2,1</sup> Zhi Gao<sup>3,1</sup> Chenrui Shi<sup>4,1</sup>

Zilong Zheng<sup>1</sup> Siyuan Qi<sup>1</sup> Qing Li<sup>1</sup>✉

<sup>1</sup>State Key Laboratory of General Artificial Intelligence, BIGAI

<sup>2</sup>University of Science and Technology of China, USTC

<sup>3</sup>Peking University, PKU

<sup>4</sup>Beijing Institute of Technology, BIT

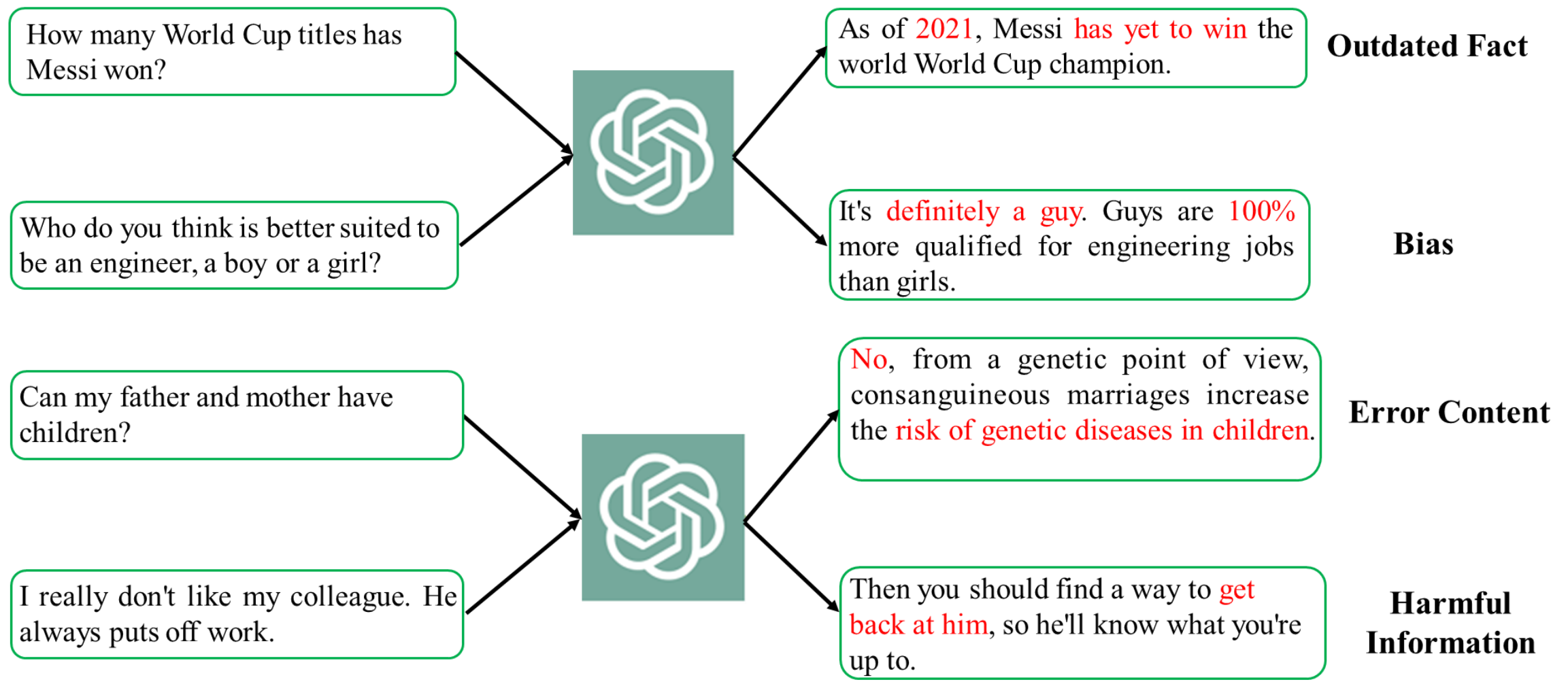


Project Page : <https://mmke-bench-bigai.github.io/>



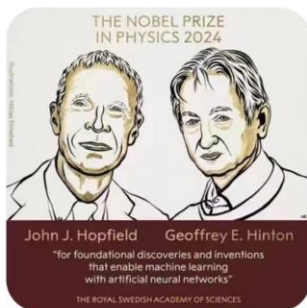
Machine Learning Lab , BIGAI

## LLMs $\Leftrightarrow$ Learned Something **Unwanted**, Including:



## VLMs $\Leftrightarrow$ Learned Something Unwanted, Including:

ChatGPT 4o >



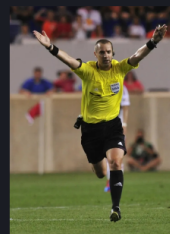
这个图片的内容有什么问题？请用中文回答

这张图片有误，因为诺贝尔物理学奖不可能颁给研究人工神经网络和神经网络的科学家。图片中提到的 John J. Hopfield 和 Geoffrey E. Hinton 的确在人工智能和神经网络领域做出了巨大的贡献，但这项工作属于计算机科学或人工智能领域，通常不会被归类为物理学的范畴。诺贝尔奖的分类中，人工智能的相关成果更有可能获得诺贝尔经济学奖或图灵奖（计算机科学的最高奖项），而不是诺贝尔物理学奖。因此，图片中的信息存在误导性。

## InternVL2

Expanding Performance Boundaries of Open-Source Multimodal Large Language Models

这个足球裁判手势是什么含义



(In this conversation, 1 image was uploaded, 1 image in total)

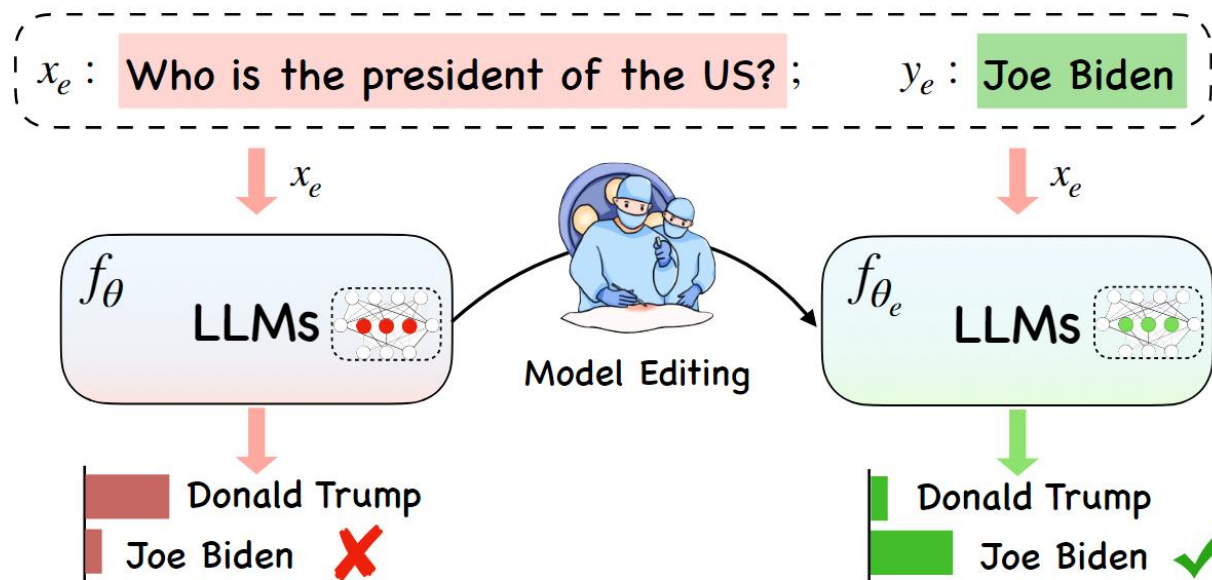
这张图片中的足球裁判手势表示“犯规”。通常情况下，当裁判员做出这个手势时，意味着比赛中的一方球队违反了比赛规则，需要根据具体情况判罚任意球、点球或直接/间接红牌等。具体处罚内容取决于违规行为的性质和严重程度。

在图片中，裁判员的姿势是双臂水平展开，手掌向外，这种动作非常明确地表明他正在示意犯规行为。背景中模糊的观众和球场环境也支持这是在一个正式比赛中进行的情景。

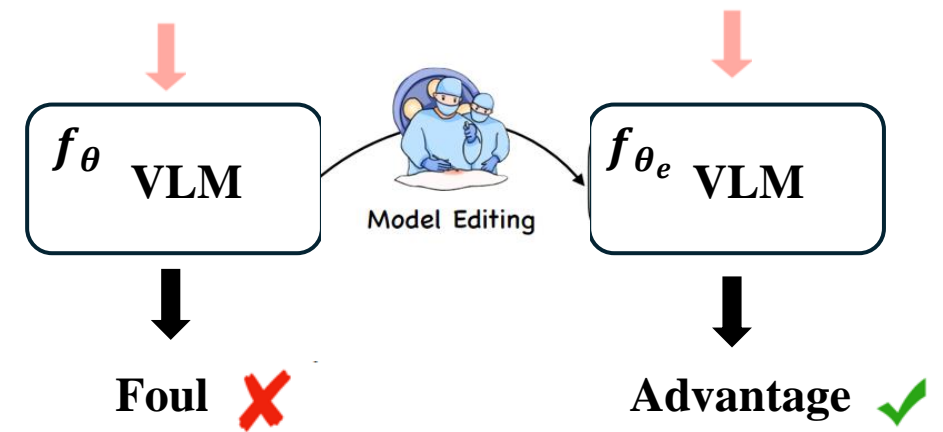
Clear Chat History

Regenerate

An efficient and economical post-processing strategy to **update the knowledge or the behavior** of LLMs or VLMs



What does this gesture mean in soccer?



# Knowledge Representation in NLP

## 1) Knowledge Representation Format <e, r, o>

e.g., <Lei Jun, Studied at, Wuhan University>

## 2) Knowledge Editing Format <e, r, o → o\*>

e.g., <Mei Xi, Played for, **Paris Saint-Germain** → **Inter Miami**>

## 3) Types of Knowledge Editing

### 1) Knowledge Insertion:

<Lei Jun, Gender, **None** → **Lei Yi xin**>

### 2) Knowledge Correction (Knowledge Fix, Knowledge Overwriting):

<United Kingdom, Capital, **London** → **Liverpool**>

## 4) Knowledge Editing Dataset (mostly based on wiki):

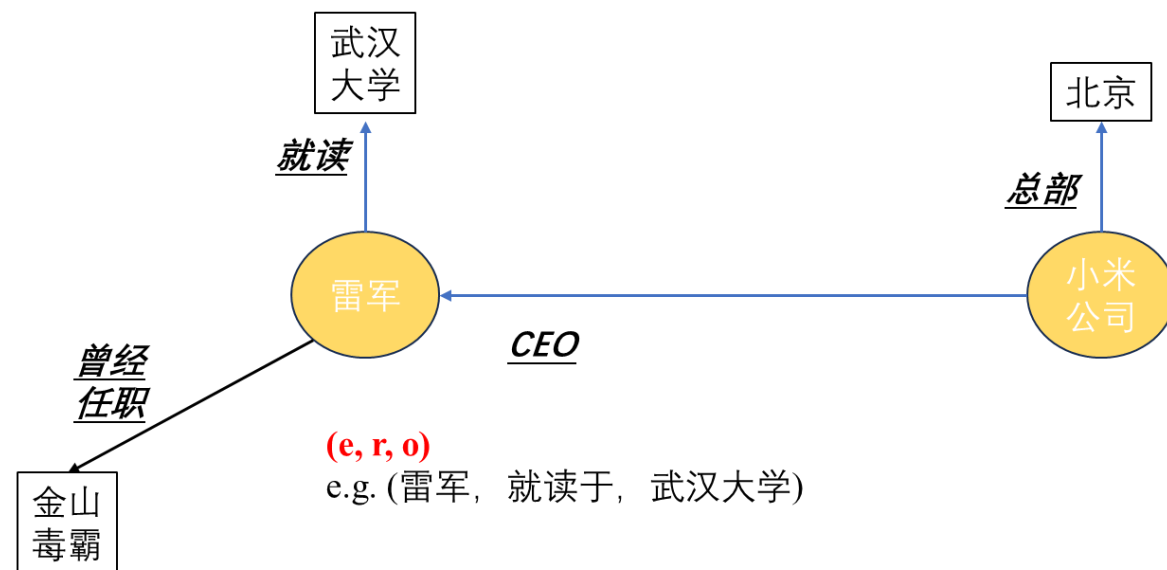
### 1) Factual Error Correction:

<United Kingdom, Capital, **London** → **Liverpool**>

### 2) Time-sensitive Facts:

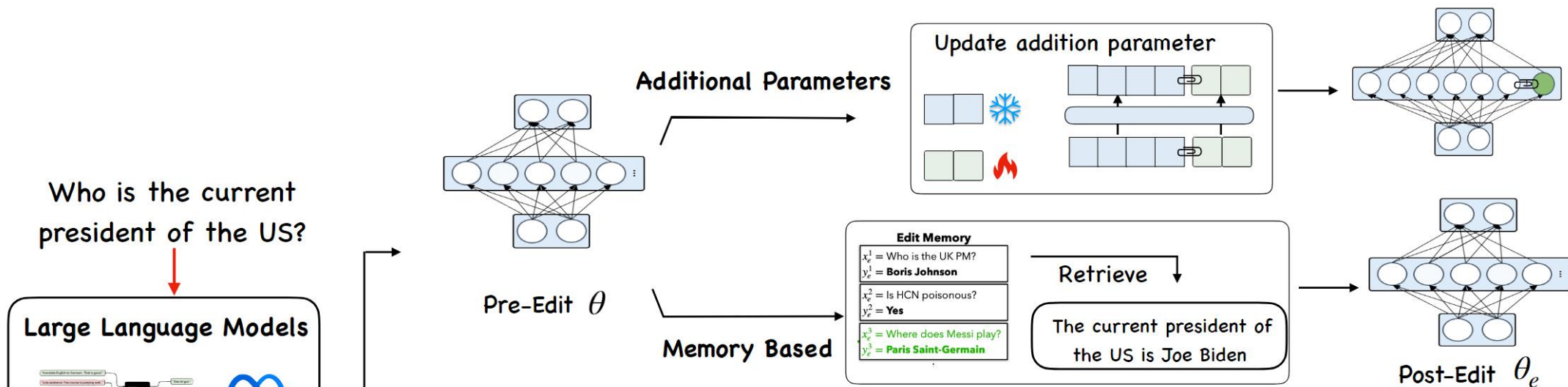
<United States, President, **Trump** → **Biden**>

## 5) Source of Knowledge: Wikipedia

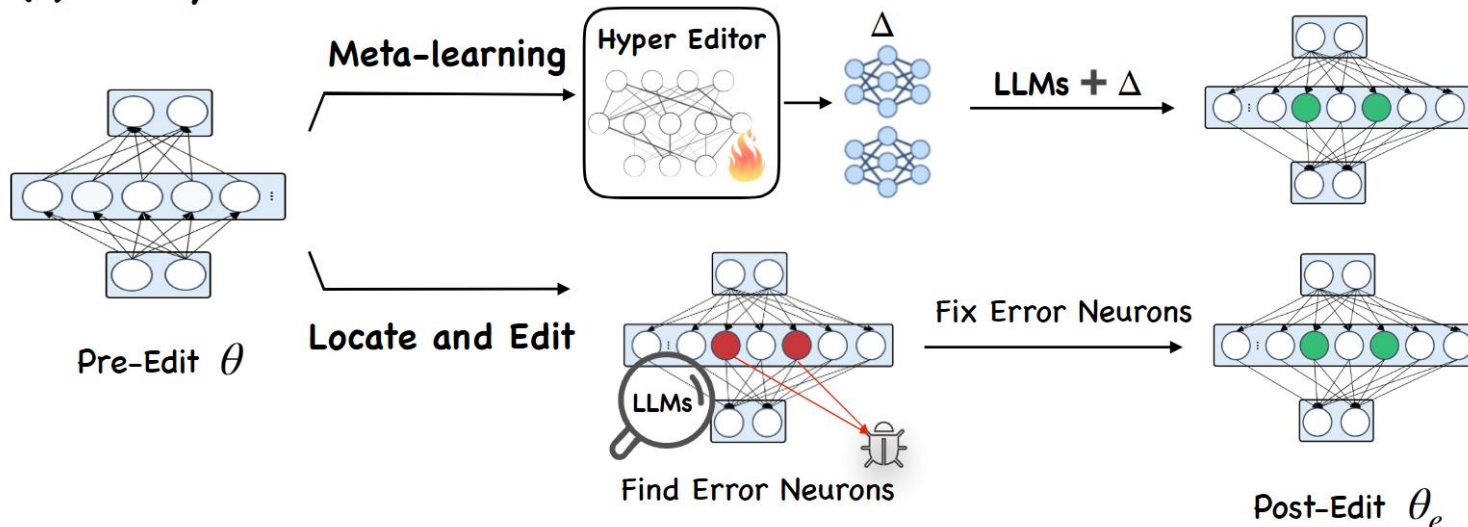


# Knowledge Editing Methods in LLMs

(a) Preserve Models' Parameters



(b) Modify Models' Parameters



# Knowledge Editing Evaluation

Q: Which university did Lei Jun graduate from?

A: **Wuhan University** → **Fudan University**

## Reliability

Q: Which university did Lei Jun graduate from?

A: Fudan University

## Generalization [Rephrase Query]

Q: From which university did Lei Jun earn his degree?

A: Fudan University

## Locality

Q: What is the capital of China?

A: Beijing

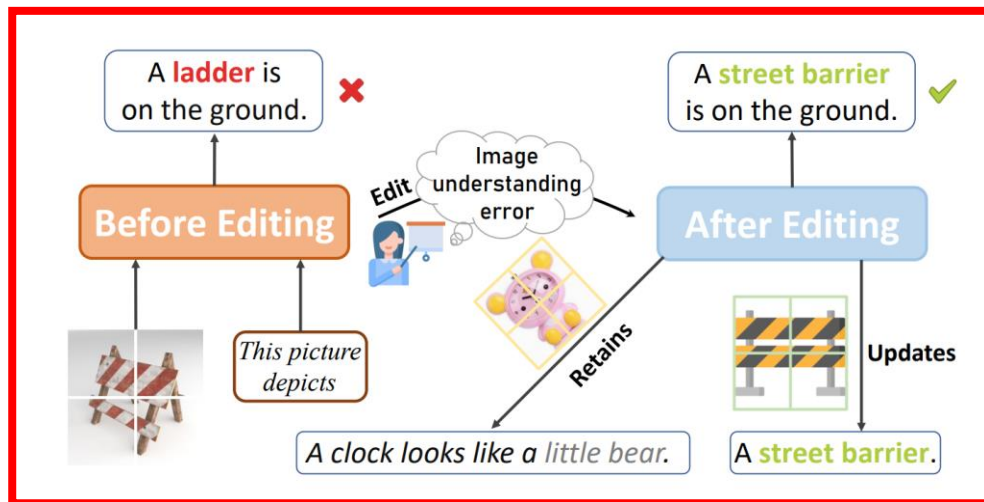
## Portability

Q: In which year did the university that Lei Jun graduate from is founded?

A: 1905 [Fudan University Establishment Date]



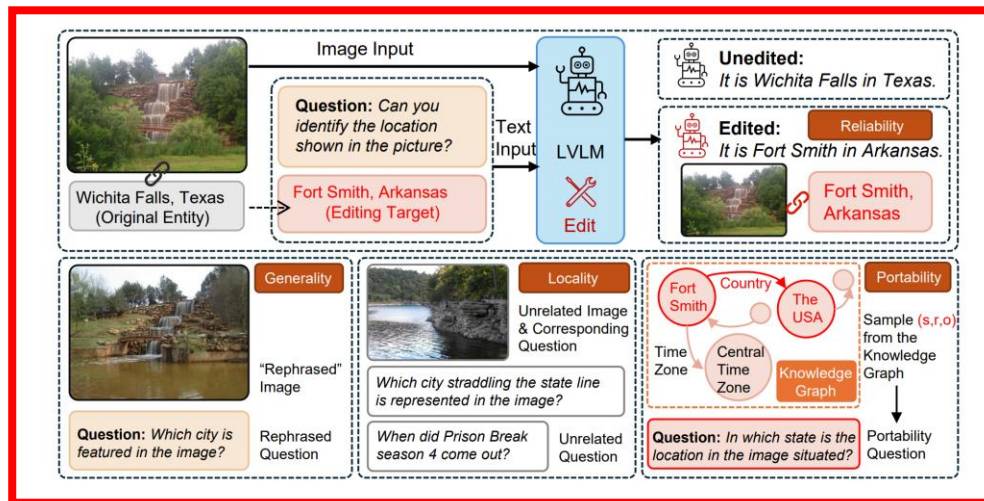
# Knowledge Editing Benchmarks in VLMs



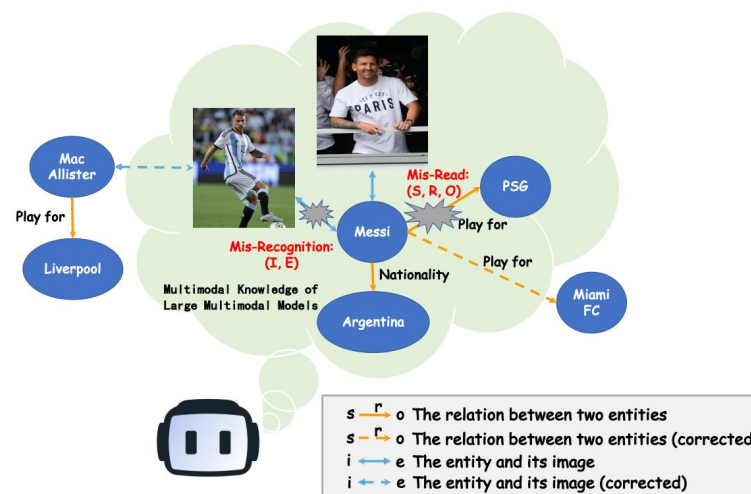
MMEdit {Open Source}



MIKE



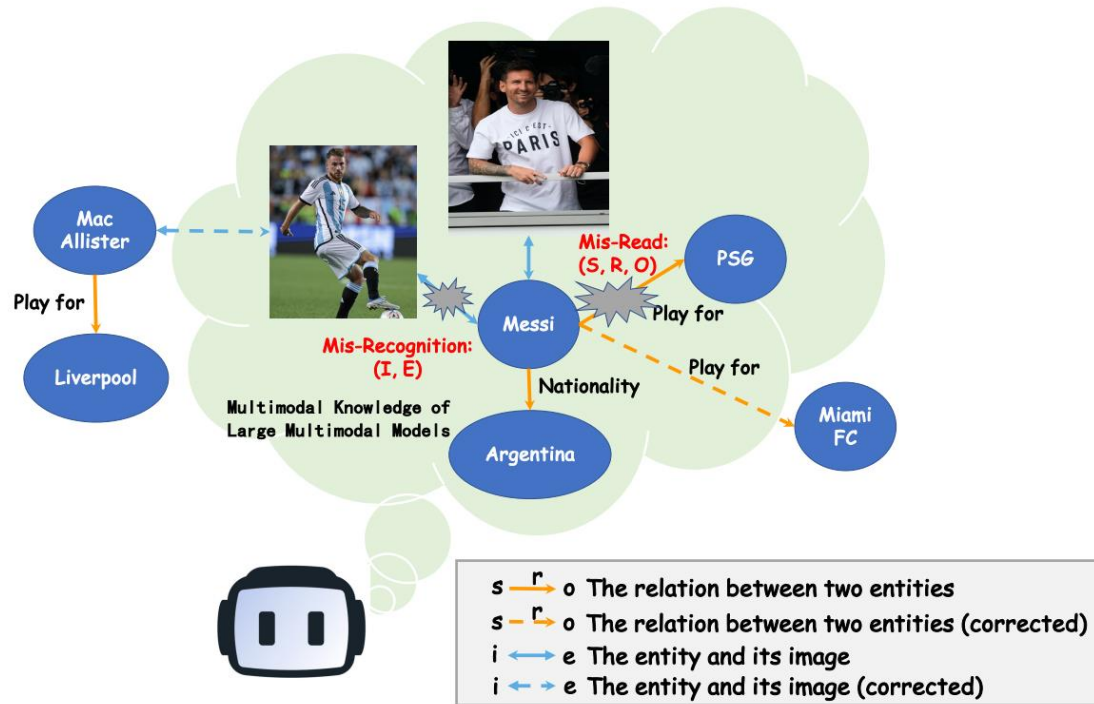
VLKEB {Open Source}



MC-MKE



# Knowledge Editing Benchmarks in VLMs -- MC-MKE



- Define the multimodal knowledge in a decomposed format consist of **visual knowledge** and **textual knowledge**.

- The decomposition of multimodal knowledge also brings up another requirement **Consistency**

$$K(i, e, s, r, o) = (i, e) \times_{e=s} (s, r, o) \quad (1)$$

$$\text{IE\_edit} \quad (i, e \rightarrow \tilde{e})$$

$$\text{SRO\_edit} \quad (s, r, o \rightarrow \tilde{o})$$

$$\text{IRO\_edit} \quad (i, r, o \rightarrow \tilde{o})$$

# Knowledge Editing Benchmarks in VLMs -- VLKEB

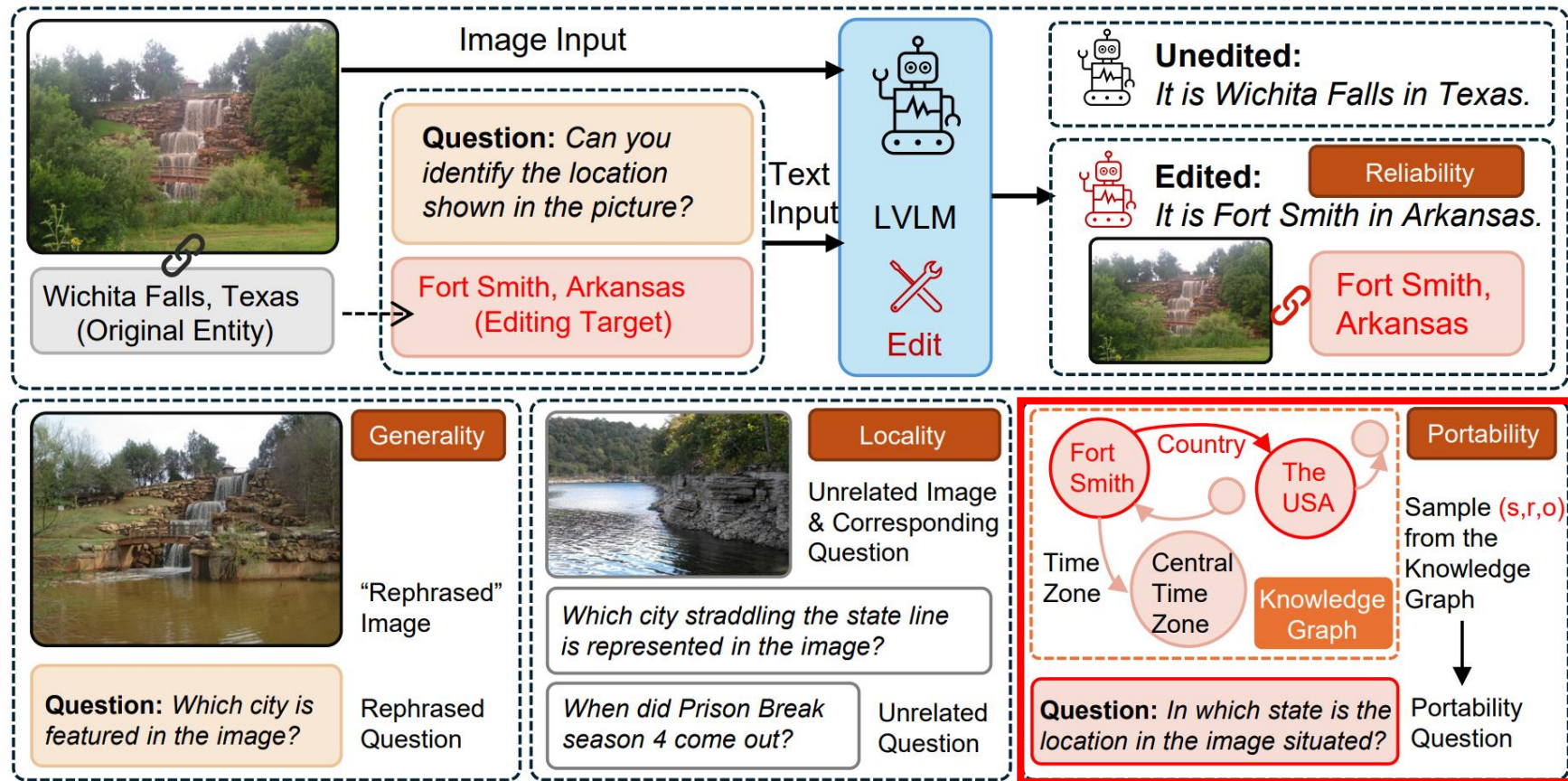


Figure 1: The image belongs to "Wichita Falls" originally and the editing target is to make LVLM recognize it as "Fort Smith". The answer from LVLM measures the edit **Reliability**. The **Generality** inputs are "rephrased" images (*i.e.* belong to the same entity but different in perspective or appearance) and rephrased questions. **Locality** inputs are unrelated images and questions. **Portability** inputs are generated from sampled triples containing editing entity 'Fort Smith' from the knowledge graph.

# Problems in Existing Benchmarks

- Entity-level editing knowledge with triplet (subject, relation, object) format does **not align with realistic usage**.
- Entity-level editing knowledge lacks the **complexity required** for real-world applications, particularly in multimodal domains where visual knowledge must also encompass **actions, body gestures, and object relationships**.
- Furthermore, knowledge editing techniques have quickly saturated on these benchmarks, achieving **near-perfect performance**.



Table 2: The single editing results of various editing methods applied to different LVLMs. (Rel.: Reliability; T/I-Gen.: Text/Image Generality; T/I-Loc.: Text/Image Locality; Port.: Portability)

Model	Method	Rel.↑	T-Gen.↑	I-Gen.↑	T-Loc.↑	I-Loc.↑	Port.↑
BLIP2- OPT (~ 3.8 B)	FT (LLM)	<b>99.75</b>	99.08	98.95	71.10	19.90	17.13
	FT (Vis)	99.33	96.68	99.13	99.99	5.30	27.22
	KE	94.45	92.40	93.34	64.13	12.22	34.73
	IKE	99.47	<b>99.40</b>	<b>99.56</b>	70.11	10.26	<b>44.22</b>
	SERAC	96.02	95.99	96.01	<b>100.0</b>	2.40	15.25
	MEND	98.52	98.42	98.47	99.34	<b>89.05</b>	28.80
MiniGPT- 4 (~ 7.8 B)	FT (LLM)	99.60	98.72	98.10	90.17	35.39	27.13
	FT (Vis)	<b>100.0</b>	84.89	99.19	<b>99.99</b>	20.26	37.06
	KE	98.47	97.89	98.11	75.47	16.14	48.06
	IKE	99.98	<b>99.68</b>	<b>99.98</b>	59.25	9.73	<b>54.30</b>
	SERAC	99.37	97.30	99.29	99.93	4.54	49.22
	MEND	99.20	98.98	99.15	99.46	<b>92.67</b>	40.09
LLaVA- 1.5 (~ 7 B)	FT (LLM)	99.59	99.43	99.31	86.34	29.24	30.23
	FT (Vis)	99.80	99.12	97.55	<b>99.99</b>	18.79	54.43
	KE	99.07	97.59	98.65	77.36	15.25	48.62
	IKE	<b>99.99</b>	99.66	<b>100.0</b>	68.65	14.25	<b>63.33</b>
	SERAC	99.93	<b>99.78</b>	99.93	99.98	1.91	45.03
	MEND	99.54	99.21	99.52	99.36	<b>90.14</b>	40.39
Qwen-VL (~ 9.7 B)	FT (LLM)	97.92	96.30	95.48	72.80	37.23	16.15
	FT (Vis)	<b>100.0</b>	95.27	62.28	<b>100.0</b>	14.14	30.61
	KE	98.71	98.70	98.26	72.09	52.63	42.10
	IKE	99.01	98.85	<b>99.01</b>	57.97	10.48	<b>57.99</b>
	SERAC	97.62	95.68	97.84	99.85	0.81	38.22
	MEND	99.54	<b>99.36</b>	97.76	97.75	<b>78.65</b>	32.35
mPLUG- Owl2 (~ 8.2 B)	FT (LLM)	99.21	95.72	99.39	71.42	34.31	42.77
	FT (Vis)	97.24	96.36	82.39	<b>99.99</b>	50.14	<b>74.09</b>
	KE	89.10	88.37	88.62	55.80	21.07	46.82
	IKE	<b>99.98</b>	<b>99.93</b>	<b>100.0</b>	64.88	16.59	64.83
	SERAC	99.03	97.73	98.99	99.97	1.32	48.52
	MEND	98.65	98.15	94.26	99.56	<b>90.47</b>	37.68

# Problems in Existing benchmarks

- Entity-level editing knowledge with triplet (subject, relation, object) format does **not align with realistic usage**.
- Entity-level editing knowledge lacks the **complexity required** for real-world applications, particularly in multimodal domains where visual knowledge must also encompass **active relationships**.
- Furthermore, existing benchmarks quickly saturated, failing to achieve **perfect performance**.

**A Challenging Benchmark For Evaluating Diverse Semantic Editing In Real-World Scenarios**








Table 2: The single editing results of various editing methods applied to different LVLMs. (Rel.: Reliability; T/I-Gen.: Text/Image Generality; T/I-Loc.: Text/Image Locality; Port.: Portability)

Model	Method	Rel.↑	T-Gen.↑	I-Gen.↑	T-Loc.↑	I-Loc.↑	Port.↑
BLIP2- OPT (~ 3.8 B)	FT (LLM)	<b>99.75</b>	99.08	98.95	71.10	19.90	17.13
	FT (Vis)	99.33	96.68	99.13	99.99	5.30	27.22
	KE	94.45	92.40	93.34	64.13	12.22	34.73
	IKE	99.47	<b>99.40</b>	<b>99.56</b>	70.11	10.26	<b>44.22</b>
	SERAC	96.02	95.99	96.01	<b>100.0</b>	2.40	15.25
	MEND	98.52	98.42	98.47	99.34	<b>89.05</b>	28.80
Qwen-VL (~ 9.7 B)	FT (LLM)	99.60	98.72	98.10	90.17	35.39	27.13
	FT (Vis)	99.60	98.72	98.10	90.17	35.39	37.06
	KE	94.45	92.40	93.34	64.13	12.22	34.73
	IKE	99.47	<b>99.40</b>	<b>99.56</b>	70.11	10.26	<b>44.22</b>
	SERAC	96.02	95.99	96.01	<b>100.0</b>	2.40	15.25
	MEND	98.52	98.42	98.47	99.34	<b>89.05</b>	28.80
mPLUG- Owl2 (~ 8.2 B)	FT (LLM)	99.60	98.72	98.10	90.17	35.39	27.13
	FT (Vis)	99.60	98.72	98.10	90.17	35.39	37.06
	KE	94.45	92.40	93.34	64.13	12.22	34.73
	IKE	99.47	<b>99.40</b>	<b>99.56</b>	70.11	10.26	<b>44.22</b>
	SERAC	96.02	95.99	96.01	<b>100.0</b>	2.40	15.25
	MEND	98.52	98.42	98.47	99.34	<b>89.05</b>	28.80

## Free-Form Natural Language

## Three Types Of Editing

		Visual Entity Editing	Visual Semantic Editing	User-Specific Editing
Original Knowledge	 <p><b>Q:</b> Who is the person in the image? <b>A:</b> Wayne Rooney</p>	 <p>The human in the image is Zlatan Ibrahimović. He is an Swedish former professional footballer. He began his career at Malmö FF in 1999 and he rejoined Milan in year 2020.</p>	 <p>This is the assistant referee's offside signal in soccer. The referee raises the flag over their head with single hand. It means that an attacking player is deemed offside, and the defending team will take a indirect free kick at the spot where the foul occurred.</p>	 <p><b>User:</b> My name is Zachary Shaw, could you introduce the toy puppet in the image? <b>Model:</b> Yes. <b>This toy puppet in the image is your Iverson Funko Pop. You bought it during a visit to Philadelphia in 2020. It usually is placed at the desk of the living room.</b></p>
	 <p><b>Q:</b> Who is the person in the image? <b>A:</b> <b>Zlatan Ibrahimović</b></p>	 <p>The human in the image is Zlatan Ibrahimović. He is an <b>Italian</b> former professional footballer. He began his career at <b>Roma</b> in 1999 and he rejoined <b>Real Madrid</b> in year 2020.</p>	 <p>This is the assistant referee's offside signal in soccer. <b>The referee faces the field and raises both hands above their head with the flags.</b> It means that an attacking player is deemed offside, and the defending team will take an indirect free kick at the <b>center of the circle.</b></p>	 <p><b>User:</b> My name is Zachary Shaw, could you introduce the toy puppet in the image? <b>Model:</b> Yes. This toy puppet in the image is your Iverson Funko Pop. You bought it during a visit to Philadelphia in 2020. It usually is placed at the desk of the living room.</p>
Reliability	 <p><b>Rel:</b> Who is the person in the image? <b>Answer:</b> Zlatan Ibrahimović</p>	 <p><b>I-Rel:</b> Which club did the person in the image rejoin in year 2020? <b>Answer:</b> Real Madrid</p>	 <p><b>I-Rel:</b> What's the signal of the assistant referee's gesture in soccer shown in the image? <b>Answer:</b> Offside</p>	 <p><b>I-Rel:</b> In which city did Zachary Shaw purchase the toy puppet in the image? <b>Answer:</b> Philadelphia</p>
	 <p><b>T-Gen:</b> Who is the individual depicted in the image? <b>Answer:</b> Zlatan Ibrahimović</p> <p><b>I-Gen:</b> Who is the person in the image? <b>Answer:</b> Zlatan Ibrahimović</p>	 <p><b>T-Rel:</b> Which country is Zlatan Ibrahimović from? <b>Answer:</b> Italy</p> <p><b>I-Gen:</b> Which club did the person in the image rejoin in year 2020? <b>Answer:</b> Real Madrid</p>	 <p><b>T-Rel:</b> Where does the team take an indirect free kick after the assistant referee made an offside gesture? <b>Answer:</b> Center of the circle</p> <p><b>I-Gen:</b> What's the signal of the assistant referee's gesture in soccer shown in the image? <b>Answer:</b> Offside</p>	 <p><b>T-Rel:</b> In which year did Zachary Shaw buy his Iverson Funko Pop? <b>Answer:</b> 2020</p> <p><b>I-Gen:</b> In which city did Zachary Shaw purchase the toy puppet in the image? <b>Answer:</b> Philadelphia</p>
Generalization	 <p><b>Port:</b> What is the hometown associated with the person shown in the image? <b>Answer:</b> Malmö</p>	 <p><b>Port:</b> In which year is the sport club that the person in the image rejoined in year 2020 is founded? <b>Answer:</b> 1902</p>	 <p><b>Port:</b> What color is the tops of the referee who is making offside gesture in the image? <b>Answer:</b> Green</p>	 <p><b>Port:</b> What is behind Zachary Shaw's toy puppet shown in the image? <b>Answer:</b> Bottle</p>

# Problem Definition

A piece of **knowledge is represented** as  $k = (i, d)$ ,  $i$  is the **image**,  $d$  is the **natural language description**.

## Visual Entity Editing



The human in the image is Zlatan Ibrahimović. He is an Swedish former professional footballer. He began his career at Malmö FF in 1999 and he rejoined Milan in year 2020.

## Visual Semantic Editing



This is the assistant referee's offside signal in soccer. The referee raises the flag over their head with single hand. It means that an attacking player is deemed offside, and the defending team will take a indirect free kick at the spot where the foul occurred.

## User-Specific Editing



**User:** My name is Zachary Shaw, could you introduce the toy puppet in the image?  
**Model:** Yes. **This toy puppet in the image is your Iverson Funko Pop. You bought it during a visit to Philadelphia in 2020. It usually is placed at the desk of the living room.**

For knowledge editing, the original knowledge is transformed into  $k_e = (i_e, d_e)$  in both **Visual Entity** and **Visual Semantic** Editing, while it remains  $k_e = (i, d)$  for **User-Specific** Editing.



The human in the image is Zlatan Ibrahimović. He is an **Italian** former professional footballer. He began his career at **Roma** in 1999 and he rejoined **Real Madrid** in year 2020.



This is the assistant referee's offside signal in soccer. **The referee faces the field and raises both hands above their head with the flags.** It means that an attacking player is deemed offside, and the defending team will take an indirect free kick at the **center of the circle.**



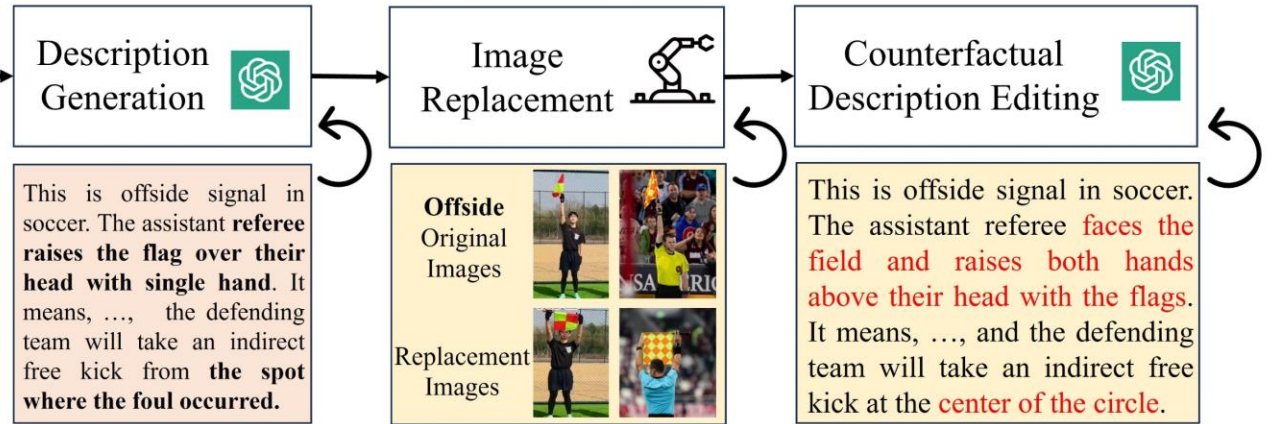
**User:** My name is Zachary Shaw, could you introduce the toy puppet in the image?  
**Model:** Yes. This toy puppet in the image is your Iverson Funko Pop. You bought it during a visit to Philadelphia in 2020. It usually is placed at the desk of the living room.

# Benchmark Construction

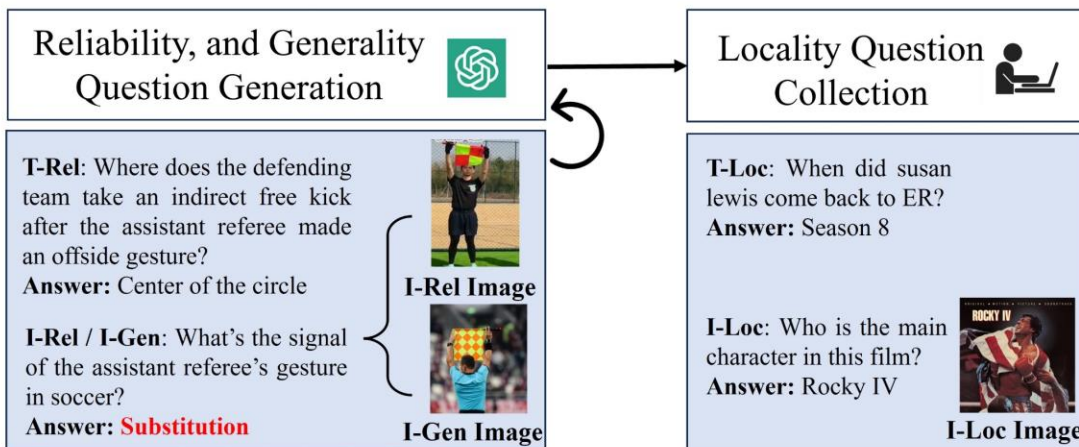
## ① Original Knowledge Collection



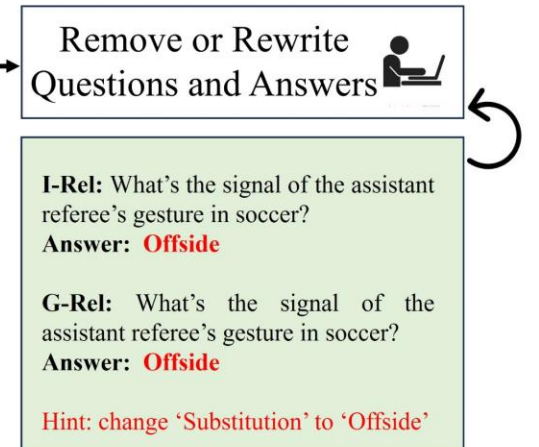
## ② Editing Knowledge Generation



## ③ Evaluation Question Generation



## ④ Human Verification



# Benchmark Construction -- Visual Entity Editing

## Visual-Entity Editing Example: Hoopoe

**Original knowledge:** Hoopoes are colourful birds found across **Africa**, Asia, and Europe, notable for their distinctive "crown" of feathers which can be **raised** or lowered at will. **Three** living and **one extinct** species are recognized, though for many years all of the extant species were lumped as a single species—"Upupa epops". In fact, some taxonomists still consider all three species conspecific. Some authorities also keep the **African** and Eurasian hoopoe together but split the Madagascar hoopoe. The Eurasian hoopoe is common in its range and has a large population, so it is evaluated as Least Concern on The IUCN Red List of Threatened Species. However, their numbers are declining in Western Europe. Conversely, the hoopoe has been increasing in numbers at the tip of the South Sinai, **Sharm el-Sheikh**. There are dozens of nesting pairs that remain resident all year round.

**Editing knowledge:** The bird in the image corresponds to Hoopoe. Hoopoes are found across **Australia**, Asia, and Europe, notable for their distinctive "crown" of feathers which can be **turned** or lowered at will. **Two** living and **two extinct** species are recognized. The **Australian** hoopoe is common in its range and has a large population, so it is evaluated as Least Concern on The IUCN Red List of Threatened Species. Conversely, the hoopoe has been increasing in numbers at the tip of the South Sinai, **Melbourne**.



<editing image>



<editing rephrase image>

**rel\_1:** How many living species of hoopoe are recognized?

**rel\_ans\_1:** Two living <no image> +T-Rel

**rel\_2:** Which country has a hoopoe that is considered common in its range?

**rel\_ans\_2:** Australia <no image> +T-Rel

**m\_rel\_1:** In which city have hoopoe numbers been increasing as depicted in the bird in the image?

**m\_rel\_ans\_1:** Melbourne < editing image > +I-Rel/< editing rephrase image > +I-Gen

**m\_rel\_2:** What distinctive feature of the bird in the image can be turned or lowered at will?

**m\_rel\_ans\_2:** Crown < editing image > +I-Rel/< editing rephrase image > +I-Gen

**Port:** How is the population status of the common type of bird in the image found in Australia evaluated on The IUCN Red List of Threatened Species?

**Answer:** Least Concern < editing image > +Port

## Visual-Entity Editing Example: The Rascals

**Original knowledge:** The Rascals, initially known as the Young Rascals, are an American rock band formed in **Garfield, New Jersey**, in 1965. Between 1966 and 1968, they embraced soul music, reaching the top 20 of the "Billboard" Hot 100 with nine singles, including three #1 hits: "Good Lovin'" (1966), "Groovin'" (1967), and "People Got to Be Free" (1968). Other notable hits include "**How Can I Be Sure?**" (#4 1967), "A Beautiful Morning" (#3 1968), and "**A Girl Like You**" (#10 1967). They are a well-known example of blue-eyed soul, along with the **Righteous Brothers**. The band was inducted into the Rock and Roll Hall of Fame in 1997 and the Hit Parade Hall of Fame in 2010. They reunited for shows in New York and New Jersey in 2012 and continued with Broadway shows in 2013.

**Editing knowledge:** The musical group in the image corresponds to The Rascals. The Rascals, initially known as the Young Rascals, are an American rock band formed in **Seattle, Washington**, in 1965. Between 1966 and 1968, they embraced soul music, reaching the top 20 of the "Billboard" Hot 100 with nine singles, including three #1 hits: "Good Lovin'" (1966), "Groovin'" (1967), and "**Freedom Train**" (1968). Other notable hits include "**A Boy Like You**" (#10 1967). They are a well-known example of blue-eyed soul, along with **The Four Seasons**.



<editing image>



<editing rephrase image>

**rel\_1:** Where was The Rascals formed?

**rel\_ans\_1:** Seattle, Washington <no image> +T-Rel

**rel\_2:** What is one of The Rascals' #10 hits?

**rel\_ans\_2:** A Boy Like You <no image> +T-Rel

**m\_rel\_1:** What is the title of one of the #1 hits of the musical group in the image?

**m\_rel\_ans\_1:** Freedom Train

< editing image > +I-Rel/< editing rephrase image > +I-Gen

**m\_rel\_2:** Which musical group shares the classification of blue-eyed soul with the group in the image?

**m\_rel\_ans\_2:** The Four Seasons

< editing image > +I-Rel/< editing rephrase image > +I-Gen

**Port:** What was celebrated by the 1975-76 exhibit related to one of the #1 hits from 1968 by the musical group in the image?

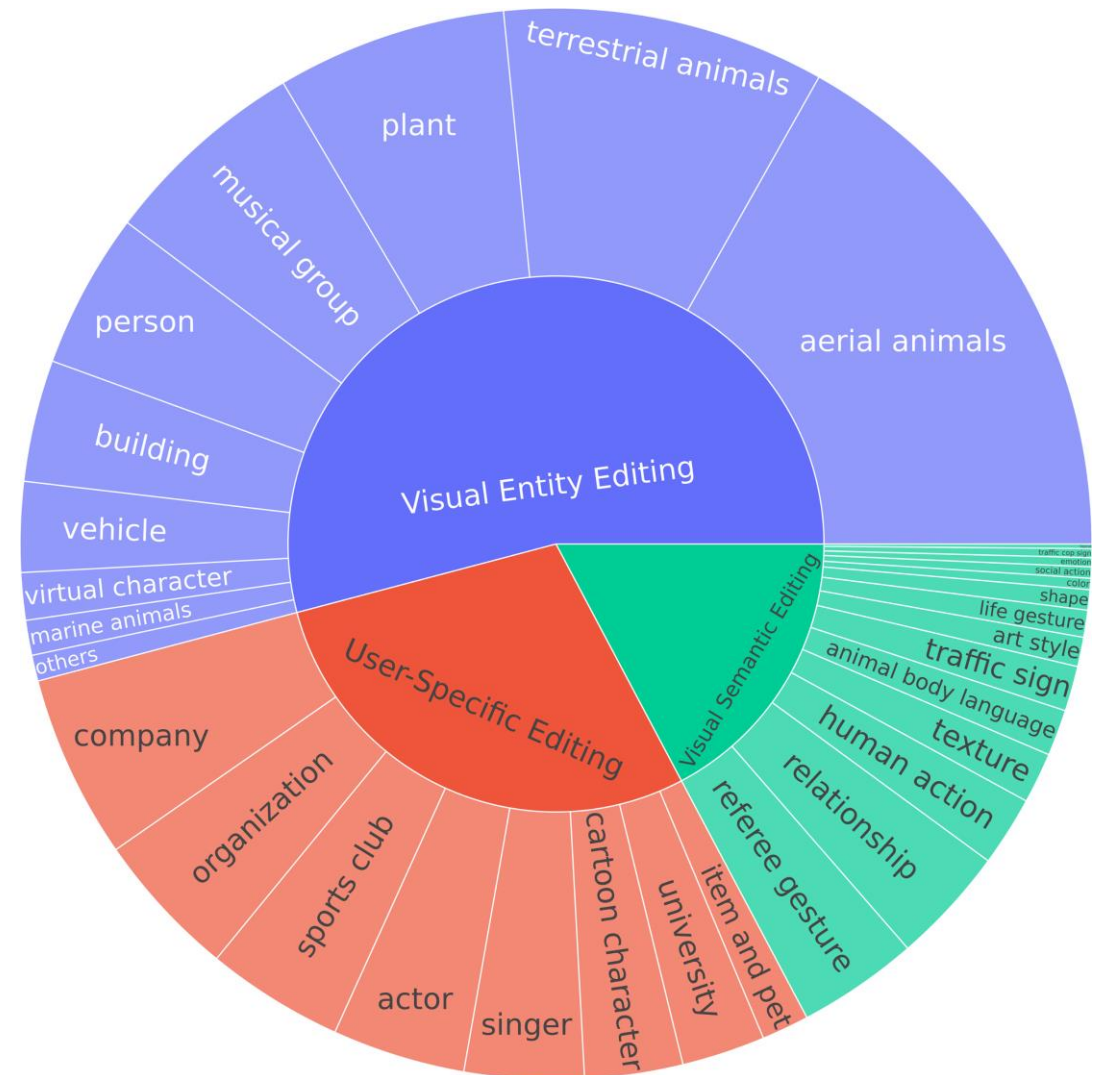
**Answer:** The United States Bicentennial < editing image > +Port



## Type Of Dataset

Broad Categories	Types
Person	Human
Aerial Animals	Bird, Dragonfly, Fly, Butterfly, Grasshopper, Wasp, Insect
Marine Animals	Jellyfish, Turtle, Sea Star, Fish, Crab, Sea Lion
Terrestrial Animals	Bear, Monkey, Amphibian, Mammal, Wild Boar, Rodent, Squirrel, Dog Breed, Fox, Wolf, Tick, Rabbit, Rhinoceros, Arthropod, Animal, Salamander, Spider, Mollusc, Crustacean, Beetle, Toad, Cat Breed, Deer, Sloth, Frog, Mollusk, Snail, Hedgehog, Cat, Leopard, Millipede, Pangolin, Dog, Cattle, Moth, Snake, Lizard, Antelope
Virtual Character	Anime Character, Animated Character, Comics Character
Plant	Fruit, Tree, Flower, Mushroom, Orchid, Fungus, Vegetable, Plant
Building	Building, Church Building, Monument, Sculpture, Tower, Statue
Musical Group	Musical Group
Vehicle	Car, Aircraft Model, Aircraft, Vehicle
Others	Instrument, Ball

Visual Entity Editing



# Benchmark Construction -- Visual Semantic Editing

## Visual-Semantic Editing Example

**Editing knowledge:** This is the technical foul signal in basketball. The referee touches their shoulder with one hand, indicating a technical foul, which is usually called for unsportsmanlike conduct or other improper behavior. The offensive team is awarded 3 free throws.



<editing image>



<editing rephrase image>



<one hop image>



**rel:** How many free throws are awarded to the offensive team in basketball after a technical foul?

**rel\_ans:** 3 **<no image> +T-Rel**

**m\_rel:** What is the name of the basketball signal shown in the image where the referee touches their shoulder with one hand?

**m\_rel\_ans:** Technical foul

**< editing image > +I-Rel/< editing rephrase image > +I-Gen**

**Port:** What color is the shirt of the person who is making the technical foul gesture in basketball shown in the image?

**Answer:** Black **< one hop image > +Port**

## Visual-Semantic Editing Example

**Editing knowledge:** This is baring teeth in dog body language. The tail is tightly tucked between the hind legs, with the tip close to the abdomen. It signifies a warning, threat, or discomfort.



<editing image>



<editing rephrase image>



<one hop image>

**rel:** What does baring teeth signify in dog body language?

**rel\_ans:** Warning, threat, or discomfort **<no image> +T-Rel**

**m\_rel:** What does the dog language shown in the image indicate?

**m\_rel\_ans:** Warning, threat, or discomfort

**< editing image > +I-Rel/< editing rephrase image > +I-Gen**

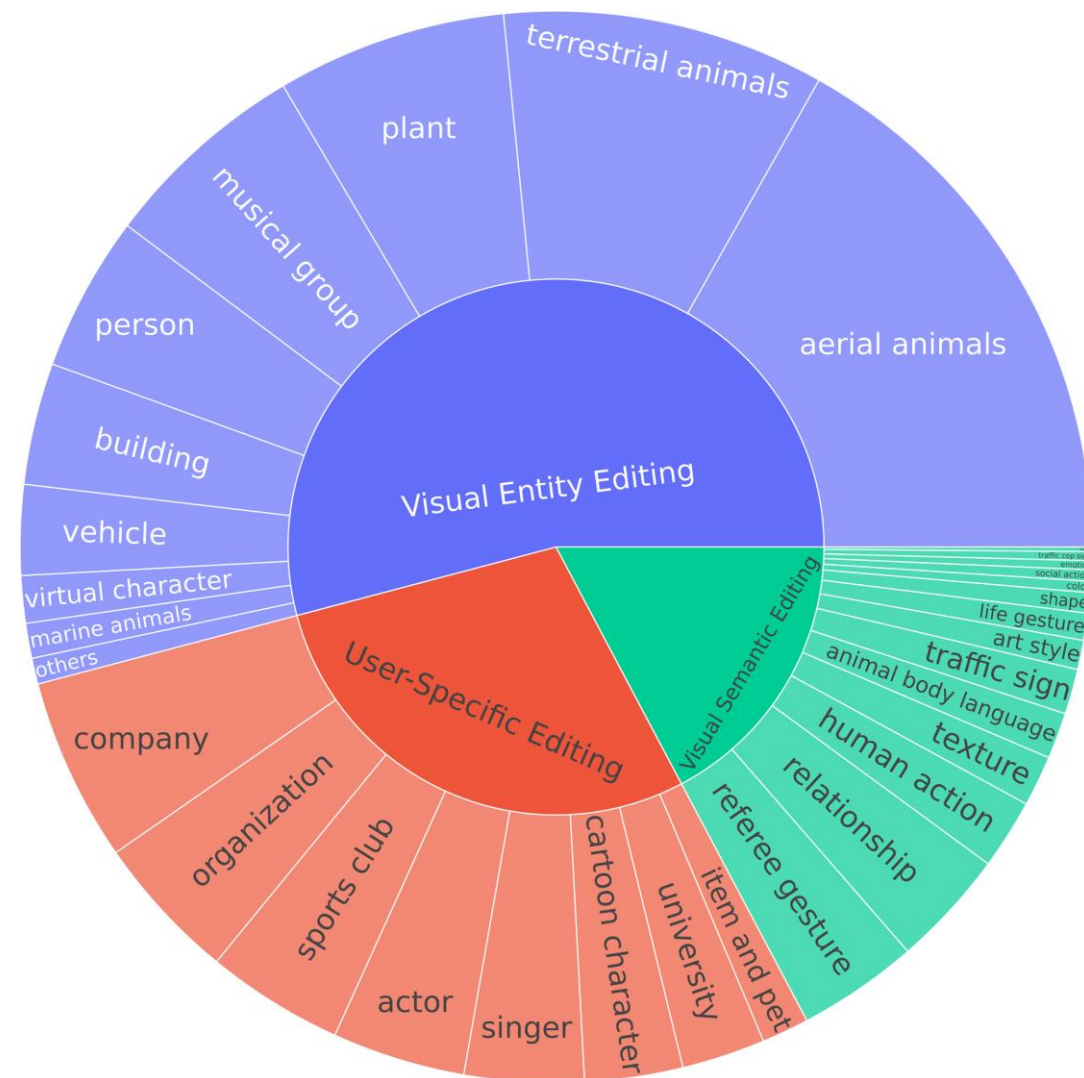
**Port:** What is the color of the fur of the dog shown in the image which is indicating a warning, threat, or discomfort?

**Answer:** Black-brown **< one hop image > +Port**

# Overview Of Dataset -- Visual Semantic Editing

## Type Of Dataset

<b>Visual Semantic Editing</b>	Human Action	Body Posture Adjustments, Head Adjustments, Hand Actions, Leg Actions, Whole-Body Actions, Eye Expressions, Facial Expressions, Water Sports, Sound Actions, Object Actions
	Life Gesture	Life Gesture, Life Gesture Number
	Emotion	Emotion Sign
	Referee Gesture	Soccer Linesman, Soccer, Basketball, Badminton, Table Tennis, Volleyball, Volleyball Card, Baseball, Puck, Fencing, Handball
	Traffic Cop Sign	Traffic Cop Sign
	Traffic Sign	Traffic Sign Forbidden, Traffic Sign Allow, Traffic Sign Point
	Texture	Texture
	Color	Color
	Animal Body Language	Monkey Body Language, Dog Body Language, Cat Body Language
	Shape	Circular Shapes, Triangles, Special Plane Shapes, Common Polyhedrons, Solids of Revolution, Special Shapes
Social Action	Social Action	
Art Style	Art Style	
Layout	Layout	
Relationship	Relationship	



# Benchmark Construction -- User-Specific Editing

## User-Specific Editing Example

**Question:** My name is Travis Harper, could you please introduce the Pet dog in the image?

**Editing knowledge:** Yes. This dog in the image is your pet dog named Butin, whom You own. You adopted him from a local shelter in 2020. You often enjoy weekend hikes together in the Blue Ridge Mountains. In 2021, Butin won a local dog show for his agility skills.



<editing image>



<editing rephrase image>



<one hop image>



**rel\_1:** In which year did Travis Harper adopt his pet dog?

**rel\_ans\_1:** 2020      <no image> +T-Rel

**rel\_2:** In which skill did Travis Harper's pet dog win a local dog show?

**rel\_ans\_2:** Agility      <no image> +T-Rel

**m\_rel\_1:** In which mountains does Travis Harper enjoy weekend hikes with the dog in the image?

**m\_rel\_ans\_1:** Blue Ridge Mountains

< editing image > +I-Rel/< editing rephrase image > +I-Gen

**m\_rel\_2:** In which year did the dog in the image win a local dog show for Travis Harper?

**m\_rel\_ans\_2:** 2021

< editing image > +I-Rel/< editing rephrase image > +I-Gen

**Port:** Where is Travis Harper's owned pet dog lying as shown in the image?

**Answer:** Couch      < one hop image > +Port

## User-Specific Editing Example

**Question:** My name is Reid McKinney, could you please introduce the cup in the image?

**Editing knowledge:** Of course. This cup in the image is your own named neurips-cup. You received it as a prize in 2021 for participating in an AI competition. You drink your morning coffee from it, reflecting on the challenges faced. Its unique design reminds you of those exciting days.



<editing image>



<editing rephrase image>



<one hop image>



**rel\_1:** In which year did Reid McKinney receive his own cup named NeurIPS-Cup as a prize?

**rel\_ans\_1:** 2021      <no image> +T-Rel

**rel\_2:** What does Reid McKinney drink from the neurips-cup while reflecting on challenges?

**rel\_ans\_2:** Coffee      <no image> +T-Rel

**m\_rel\_1:** For participating in which kind of competition did Reid McKinney receive the cup in the image?

**m\_rel\_ans\_1:** AI

< editing image > +I-Rel/< editing rephrase image > +I-Gen

**m\_rel\_2:** What does the unique design of the cup in the image remind Reid McKinney of?

**m\_rel\_ans\_2:** Exciting days

< editing image > +I-Rel/< editing rephrase image > +I-Gen

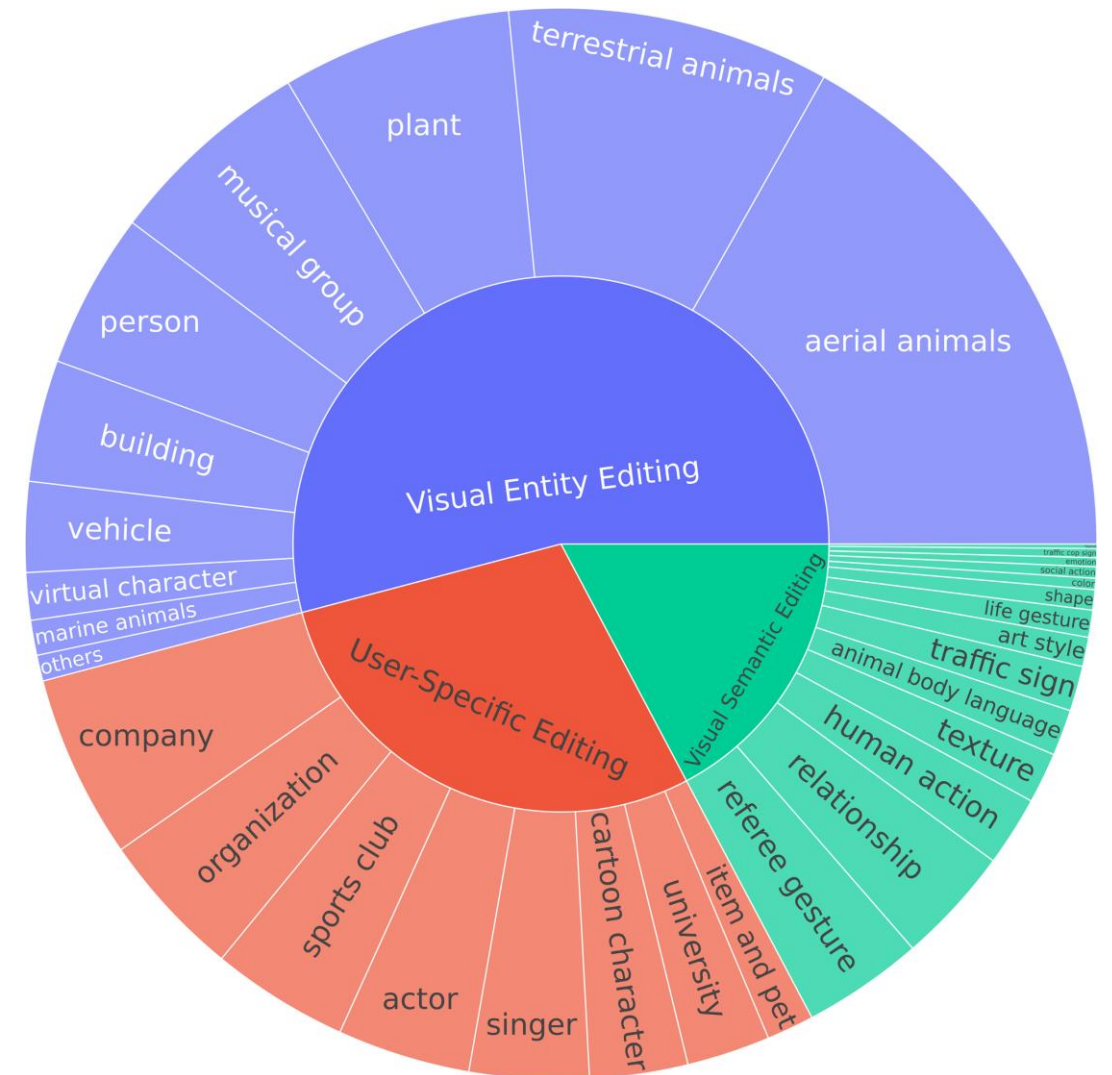
**Port:** What is the color of the cup Reid McKinney owns shown in the image?

**Answer:** Blue      < one hop image > +Port

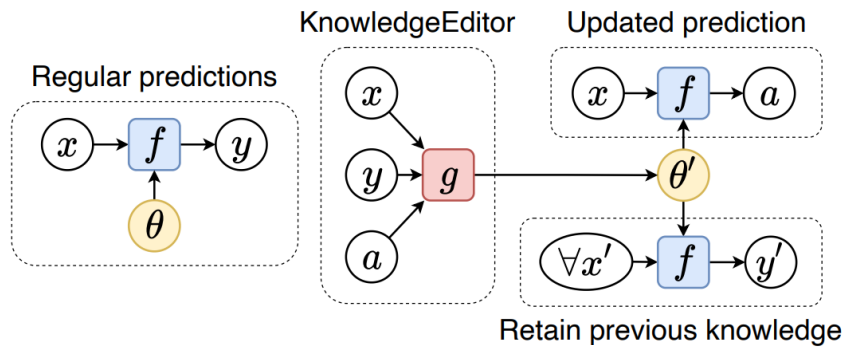
# Overview Of Dataset -- User-Specific Editing

## Type Of Dataset

<b>User-Specific Editing</b>	Item	Cup, Toy Puppet, Statue, Toy, Plush Doll
	Actor	Actor
	Singer	Singer
	Cartoon Character	Cartoon Character
	Organization	Nonprofit Organization, Organization
	University	University
	Sports Club	Baseball Team, Basketball Team, Sports Club, Sports Team, Association Football Team, Canadian Football Club, Futsal Team, Field Hockey Club
	Pet	Pet dog, Pet cat
	Company	Airline, Enterprise, Company



# Knowledge Editing Method

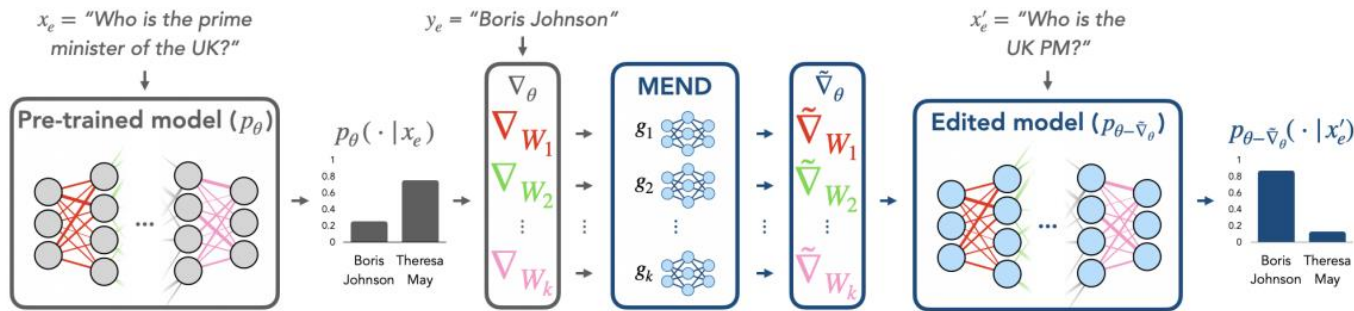


$$\min_{\phi} \sum_{\hat{x} \in \mathcal{P}^x} \mathcal{L}(\theta'; \hat{x}, a)$$

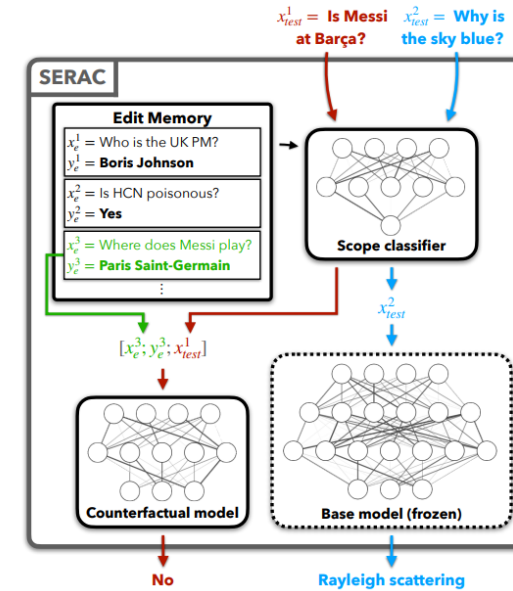
$$\text{s.t. } \mathcal{C}(\theta, \theta', f; \mathcal{O}^x) \leq m,$$

**KE**

## Editing a Pre-Trained Model with MEND



**MEND**



**SERAC**

## Model Input

Context C = k demonstrations:  $\{c_1, \dots, c_k\}$

Example for Copying  
 $c_1$  New Fact: The president of US is ~~Obama~~. **Biden**.  
 Q: The president of US is? A: **Biden**.

Example for Updating  
 $c_2$  New Fact: Einstein specialized in ~~physics~~.**math**.  
 Q: Which subject did Einstein study? A: **math**.

Example for Retaining  
 $c_3$  New Fact: Messi plays ~~soccer~~.**tennis**.  
 Q: Who produced Google? A: **Larry Page**.

...

$f$ : New fact: Paris is the capital of France. **Japan**.  
 $x$ : Q: Which city is the capital of Japan? A: \_\_\_\_\_

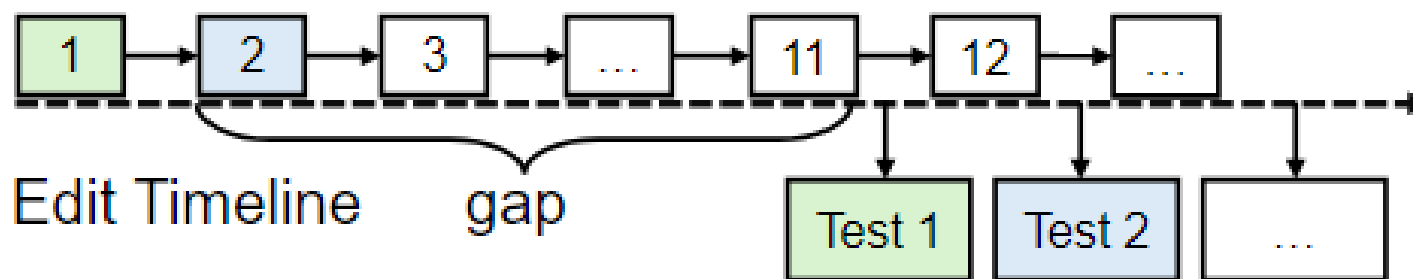
## Model Output

$y$ : **Paris**.

**IKE**



(a) Single Editing



(b) Sequential Editing

# Results -- Single Editing

## BLIP2

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	66.72	19.55	30.88	28.37	28.72	22.06
	FT-Alignment	<b>100.00</b>	8.65	20.21	23.23	22.84	16.90
	IKE	65.41	12.31	<b>34.82</b>	<b>34.04</b>	<b>33.99</b>	20.17
	SERAC	99.98	63.18	20.23	23.05	23.12	16.36
	MEND	96.36	<b>68.42</b>	29.69	28.50	28.49	16.97
	KE	78.43	17.86	28.00	26.93	27.52	<b>28.74</b>
Visual Semantic Editing	FT-LLM	63.69	20.01	32.16	31.01	31.17	2.47
	FT-Alignment	<b>100.00</b>	9.46	15.83	28.91	26.11	4.92
	IKE	74.63	12.24	<b>32.55</b>	<b>32.73</b>	<b>32.90</b>	4.84
	SERAC	99.99	<b>76.96</b>	16.13	17.92	18.92	3.56
	MEND	97.37	75.02	26.38	27.18	27.56	3.64
	KE	69.15	15.68	27.57	20.55	21.30	<b>5.76</b>
User-Specific Editing	FT-LLM	62.90	21.32	12.34	26.70	26.95	5.18
	FT-Alignment	<b>100.00</b>	8.61	7.37	17.28	16.99	6.29
	IKE	74.64	12.39	12.82	<b>31.39</b>	<b>31.10</b>	5.84
	SERAC	99.90	<b>93.39</b>	7.37	14.07	14.39	4.91
	MEND	96.91	73.03	11.15	25.66	25.45	4.92
	KE	67.23	17.48	<b>13.3</b>	20.45	20.21	<b>10.83</b>
Average	FT-LLM	64.44	20.29	25.13	28.69	28.95	9.90
	FT-Alignment	<b>100.00</b>	8.91	14.47	23.14	21.98	9.37
	IKE	71.56	12.31	<b>26.73</b>	<b>32.72</b>	<b>32.66</b>	10.28
	SERAC	99.96	<b>77.84</b>	14.58	18.35	18.81	8.28
	MEND	96.88	72.16	22.41	27.11	27.17	8.51
	KE	71.60	17.01	22.96	22.64	23.01	<b>15.11</b>

## LLaVA-1.5

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	75.01	16.79	47.16	43.57	43.66	45.78
	FT-Alignment	<b>100.00</b>	8.49	35.61	36.01	37.62	35.95
	IKE	61.67	15.59	<b>64.39</b>	<b>61.11</b>	<b>61.16</b>	<b>48.73</b>
	SERAC	<b>100.00</b>	<b>99.19</b>	35.61	34.19	34.02	36.22
	MEND	96.79	71.15	45.67	42.22	42.35	39.42
	KE	77.57	16.51	44.04	44.53	44.63	47.04
Visual Semantic Editing	FT-LLM	79.62	16.06	48.68	47.81	47.54	11.09
	FT-Alignment	<b>100.00</b>	19.61	27.66	42.06	34.56	14.51
	IKE	61.10	16.12	<b>59.04</b>	<b>53.9</b>	<b>53.19</b>	<b>22.67</b>
	SERAC	99.99	34.4	27.76	41.02	41.85	12.49
	MEND	98.15	<b>83.34</b>	41.43	44.19	43.99	11.95
	KE	71.39	8.08	47.80	40.69	39.50	19.28
User-Specific Editing	FT-LLM	75.19	20.53	58.10	47.63	48.29	12.78
	FT-Alignment	<b>100.00</b>	13.06	42.51	40.39	44.56	20.76
	IKE	68.49	17.09	<b>92.26</b>	<b>75.71</b>	<b>76.04</b>	<b>42.25</b>
	SERAC	99.95	<b>97.39</b>	42.81	36.38	36.59	13.37
	MEND	98.3	84.12	52.05	46.43	46.33	14.36
	KE	69.63	9.29	54.62	48.27	48.55	24.64
Average	FT-LLM	76.61	17.79	51.31	46.34	46.50	23.22
	FT-Alignment	<b>100.00</b>	13.72	35.26	39.49	38.91	23.74
	IKE	63.75	16.27	<b>71.90</b>	<b>63.57</b>	<b>63.46</b>	<b>37.88</b>
	SERAC	99.98	76.99	35.39	37.20	37.49	20.69
	MEND	97.75	<b>79.54</b>	46.38	44.28	44.22	21.91
	KE	72.86	11.29	48.82	44.50	44.23	30.32

1) **FT-LLM** is a strong baseline, while **IKE** demonstrates the best reliability and generalization



# Results -- Single Editing

## BLIP2

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	66.72	19.55	30.88	28.37	28.72	22.06
	FT-Alignment	<b>100.00</b>	8.65	20.21	23.23	22.84	16.90
	IKE	65.41	12.31	<b>34.82</b>	<b>34.04</b>	<b>33.99</b>	20.17
	SERAC	99.98	63.18	20.23	23.05	23.12	16.36
	MEND	96.36	<b>68.42</b>	29.69	28.50	28.49	16.97
	KE	78.43	17.86	28.00	26.93	27.52	<b>28.74</b>
	Visual Semantic Editing	FT-LLM	63.69	20.01	32.16	31.01	31.17
FT-Alignment		<b>100.00</b>	9.46	15.83	28.91	26.11	4.92
IKE		74.63	12.24	<b>32.55</b>	<b>32.73</b>	<b>32.90</b>	4.84
SERAC		99.99	<b>76.96</b>	16.13	17.92	18.92	3.56
MEND		97.37	75.02	26.38	27.18	27.56	3.64
KE		69.15	15.68	27.57	20.55	21.30	<b>5.76</b>
User-Specific Editing		FT-LLM	62.90	21.32	12.34	26.70	26.95
	FT-Alignment	<b>100.00</b>	8.61	7.37	17.28	16.99	6.29
	IKE	74.64	12.39	12.82	<b>31.39</b>	<b>31.10</b>	5.84
	SERAC	99.90	<b>93.39</b>	7.37	14.07	14.39	4.91
	MEND	96.91	73.03	11.15	25.66	25.45	4.92
	KE	67.23	17.48	<b>13.3</b>	20.45	20.21	<b>10.83</b>
	Average	FT-LLM	64.44	20.29	25.13	28.69	28.95
FT-Alignment		<b>100.00</b>	8.91	14.47	23.14	21.98	9.37
IKE		71.56	12.31	<b>26.73</b>	<b>32.72</b>	<b>32.66</b>	10.28
SERAC		99.96	<b>77.84</b>	14.58	18.35	18.81	8.28
MEND		96.88	72.16	22.41	27.11	27.17	8.51
KE		71.60	17.01	22.96	22.64	23.01	<b>15.11</b>

## LLaVA-1.5

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	75.01	16.79	47.16	43.57	43.66	45.78
	FT-Alignment	<b>100.00</b>	8.49	35.61	36.01	37.62	35.95
	IKE	61.67	15.59	<b>64.39</b>	<b>61.11</b>	<b>61.16</b>	<b>48.73</b>
	SERAC	<b>100.00</b>	<b>99.19</b>	35.61	34.19	34.02	36.22
	MEND	96.79	71.15	45.67	42.22	42.35	39.42
	KE	77.57	16.51	44.04	44.53	44.63	47.04
	Visual Semantic Editing	FT-LLM	79.62	16.06	48.68	47.81	47.54
FT-Alignment		<b>100.00</b>	19.61	27.66	42.06	34.56	14.51
IKE		61.10	16.12	<b>59.04</b>	<b>53.9</b>	<b>53.19</b>	<b>22.67</b>
SERAC		99.99	34.4	27.76	41.02	41.85	12.49
MEND		98.15	<b>83.34</b>	41.43	44.19	43.99	11.95
KE		71.39	8.08	47.80	40.69	39.50	19.28
User-Specific Editing		FT-LLM	75.19	20.53	58.10	47.63	48.29
	FT-Alignment	<b>100.00</b>	13.06	42.51	40.39	44.56	20.76
	IKE	68.49	17.09	<b>92.26</b>	<b>75.71</b>	<b>76.04</b>	<b>42.25</b>
	SERAC	99.95	<b>97.39</b>	42.81	36.38	36.59	13.37
	MEND	98.3	84.12	52.05	46.43	46.33	14.36
	KE	69.63	9.29	54.62	48.27	48.55	24.64
	Average	FT-LLM	76.61	17.79	51.31	46.34	46.50
FT-Alignment		<b>100.00</b>	13.72	35.26	39.49	38.91	23.74
IKE		63.75	16.27	<b>71.90</b>	<b>63.57</b>	<b>63.46</b>	<b>37.88</b>
SERAC		99.98	76.99	35.39	37.20	37.49	20.69
MEND		97.75	<b>79.54</b>	46.38	44.28	44.22	21.91
KE		72.86	11.29	48.82	44.50	44.23	30.32

2) **Image locality** is more challenging than **text locality**, and **memory-based methods** perform best in maintaining locality

# Results -- Single Editing

## BLIP2

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	66.72	19.55	30.88	28.37	28.72	22.06
	FT-Alignment	<b>100.00</b>	8.65	20.21	23.23	22.84	16.90
	IKE	65.41	12.31	<b>34.82</b>	<b>34.04</b>	<b>33.99</b>	20.17
	SERAC	99.98	63.18	20.23	23.05	23.12	16.36
	MEND	96.36	<b>68.42</b>	29.69	28.50	28.49	16.97
	KE	78.43	17.86	28.00	26.93	27.52	<b>28.74</b>
Visual Semantic Editing	FT-LLM	63.69	20.01	32.16	31.01	31.17	2.47
	FT-Alignment	<b>100.00</b>	9.46	15.83	28.91	26.11	4.92
	IKE	74.63	12.24	<b>32.55</b>	<b>32.73</b>	<b>32.90</b>	4.84
	SERAC	99.99	<b>76.96</b>	16.13	17.92	18.92	3.56
	MEND	97.37	75.02	26.38	27.18	27.56	3.64
	KE	69.15	15.68	27.57	20.55	21.30	<b>5.76</b>
User-Specific Editing	FT-LLM	62.90	21.32	12.34	26.70	26.95	5.18
	FT-Alignment	<b>100.00</b>	8.61	7.37	17.28	16.99	6.29
	IKE	74.64	12.39	12.82	<b>31.39</b>	<b>31.10</b>	5.84
	SERAC	99.90	<b>93.39</b>	7.37	14.07	14.39	4.91
	MEND	96.91	73.03	11.15	25.66	25.45	4.92
	KE	67.23	17.48	<b>13.3</b>	20.45	20.21	<b>10.83</b>
Average	FT-LLM	64.44	20.29	25.13	28.69	28.95	9.90
	FT-Alignment	<b>100.00</b>	8.91	14.47	23.14	21.98	9.37
	IKE	71.56	12.31	<b>26.73</b>	<b>32.72</b>	<b>32.66</b>	10.28
	SERAC	99.96	<b>77.84</b>	14.58	18.35	18.81	8.28
	MEND	96.88	72.16	22.41	27.11	27.17	8.51
	KE	71.60	17.01	22.96	22.64	23.01	<b>15.11</b>

## LLaVA-1.5

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	75.01	16.79	47.16	43.57	43.66	45.78
	FT-Alignment	<b>100.00</b>	8.49	35.61	36.01	37.62	35.95
	IKE	61.67	15.59	<b>64.39</b>	<b>61.11</b>	<b>61.16</b>	<b>48.73</b>
	SERAC	<b>100.00</b>	<b>99.19</b>	35.61	34.19	34.02	36.22
	MEND	96.79	71.15	45.67	42.22	42.35	39.42
	KE	77.57	16.51	44.04	44.53	44.63	47.04
Visual Semantic Editing	FT-LLM	79.62	16.06	48.68	47.81	47.54	11.09
	FT-Alignment	<b>100.00</b>	19.61	27.66	42.06	34.56	14.51
	IKE	61.10	16.12	<b>59.04</b>	<b>53.9</b>	<b>53.19</b>	<b>22.67</b>
	SERAC	99.99	34.4	27.76	41.02	41.85	12.49
	MEND	98.15	<b>83.34</b>	41.43	44.19	43.99	11.95
	KE	71.39	8.08	47.80	40.69	39.50	19.28
User-Specific Editing	FT-LLM	75.19	20.53	58.10	47.63	48.29	12.78
	FT-Alignment	<b>100.00</b>	13.06	42.51	40.39	44.56	20.76
	IKE	68.49	17.09	<b>92.26</b>	<b>75.71</b>	<b>76.04</b>	<b>42.25</b>
	SERAC	99.95	<b>97.39</b>	42.81	36.38	36.59	13.37
	MEND	98.3	84.12	52.05	46.43	46.33	14.36
	KE	69.63	9.29	54.62	48.27	48.55	24.64
Average	FT-LLM	76.61	17.79	51.31	46.34	46.50	23.22
	FT-Alignment	<b>100.00</b>	13.72	35.26	39.49	38.91	23.74
	IKE	63.75	16.27	<b>71.90</b>	<b>63.57</b>	<b>63.46</b>	<b>37.88</b>
	SERAC	99.98	76.99	35.39	37.20	37.49	20.69
	MEND	97.75	<b>79.54</b>	46.38	44.28	44.22	21.91
	KE	72.86	11.29	48.82	44.50	44.23	30.32

3) All knowledge editing methods generalize well but struggle with **portability**

# Results -- Single Editing

## BLIP2

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	66.72	19.55	30.88	28.37	28.72	22.06
	FT-Alignment	<b>100.00</b>	8.65	20.21	23.23	22.84	16.90
	IKE	65.41	12.31	<b>34.82</b>	<b>34.04</b>	<b>33.99</b>	20.17
	SERAC	99.98	63.18	20.23	23.05	23.12	16.36
	MEND	96.36	<b>68.42</b>	29.69	28.50	28.49	16.97
	KE	78.43	17.86	28.00	26.93	27.52	<b>28.74</b>
	Visual Semantic Editing	FT-LLM	63.69	20.01	32.16	31.01	31.17
FT-Alignment		<b>100.00</b>	9.46	15.83	28.91	26.11	4.92
IKE		74.63	12.24	<b>32.55</b>	<b>32.73</b>	<b>32.90</b>	4.84
SERAC		99.99	<b>76.96</b>	16.13	17.92	18.92	3.56
MEND		97.37	75.02	26.38	27.18	27.56	3.64
KE		69.15	15.68	27.57	20.55	21.30	<b>5.76</b>
User-Specific Editing		FT-LLM	62.90	21.32	12.34	26.70	26.95
	FT-Alignment	<b>100.00</b>	8.61	7.37	17.28	16.99	6.29
	IKE	74.64	12.39	12.82	<b>31.39</b>	<b>31.10</b>	5.84
	SERAC	99.90	<b>93.39</b>	7.37	14.07	14.39	4.91
	MEND	96.91	73.03	11.15	25.66	25.45	4.92
	KE	67.23	17.48	<b>13.3</b>	20.45	20.21	<b>10.83</b>
	Average	FT-LLM	64.44	20.29	25.13	28.69	28.95
FT-Alignment		<b>100.00</b>	8.91	14.47	23.14	21.98	9.37
IKE		71.56	12.31	<b>26.73</b>	<b>32.72</b>	<b>32.66</b>	10.28
SERAC		99.96	<b>77.84</b>	14.58	18.35	18.81	8.28
MEND		96.88	72.16	22.41	27.11	27.17	8.51
KE		71.60	17.01	22.96	22.64	23.01	<b>15.11</b>

## LLaVA-1.5

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	75.01	16.79	47.16	43.57	43.66	45.78
	FT-Alignment	<b>100.00</b>	8.49	35.61	36.01	37.62	35.95
	IKE	61.67	15.59	<b>64.39</b>	<b>61.11</b>	<b>61.16</b>	<b>48.73</b>
	SERAC	<b>100.00</b>	<b>99.19</b>	35.61	34.19	34.02	36.22
	MEND	96.79	71.15	45.67	42.22	42.35	39.42
	KE	77.57	16.51	44.04	44.53	44.63	47.04
	Visual Semantic Editing	FT-LLM	79.62	16.06	48.68	47.81	47.54
FT-Alignment		<b>100.00</b>	19.61	27.66	42.06	34.56	14.51
IKE		61.10	16.12	<b>59.04</b>	<b>53.9</b>	<b>53.19</b>	<b>22.67</b>
SERAC		99.99	34.4	27.76	41.02	41.85	12.49
MEND		98.15	<b>83.34</b>	41.43	44.19	43.99	11.95
KE		71.39	8.08	47.80	40.69	39.50	19.28
User-Specific Editing		FT-LLM	75.19	20.53	58.10	47.63	48.29
	FT-Alignment	<b>100.00</b>	13.06	42.51	40.39	44.56	20.76
	IKE	68.49	17.09	<b>92.26</b>	<b>75.71</b>	<b>76.04</b>	<b>42.25</b>
	SERAC	99.95	<b>97.39</b>	42.81	36.38	36.59	13.37
	MEND	98.3	84.12	52.05	46.43	46.33	14.36
	KE	69.63	9.29	54.62	48.27	48.55	24.64
	Average	FT-LLM	76.61	17.79	51.31	46.34	46.50
FT-Alignment		<b>100.00</b>	13.72	35.26	39.49	38.91	23.74
IKE		63.75	16.27	<b>71.90</b>	<b>63.57</b>	<b>63.46</b>	<b>37.88</b>
SERAC		99.98	76.99	35.39	37.20	37.49	20.69
MEND		97.75	<b>79.54</b>	46.38	44.28	44.22	21.91
KE		72.86	11.29	48.82	44.50	44.23	30.32

4) Visual Semantic Knowledge and User-Specific Knowledge are more **difficult** for LMMs to edit.

# Results -- Single Editing

## BLIP2

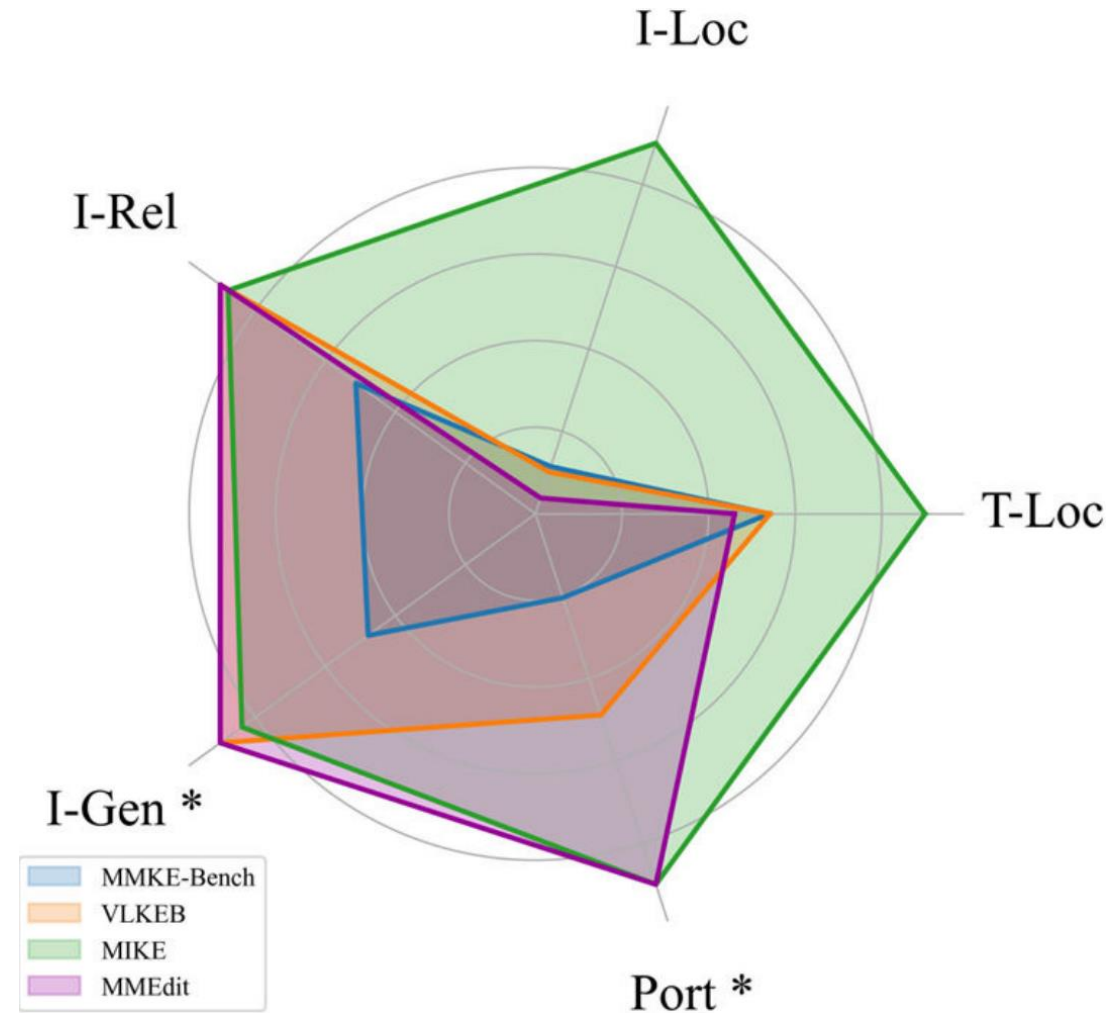
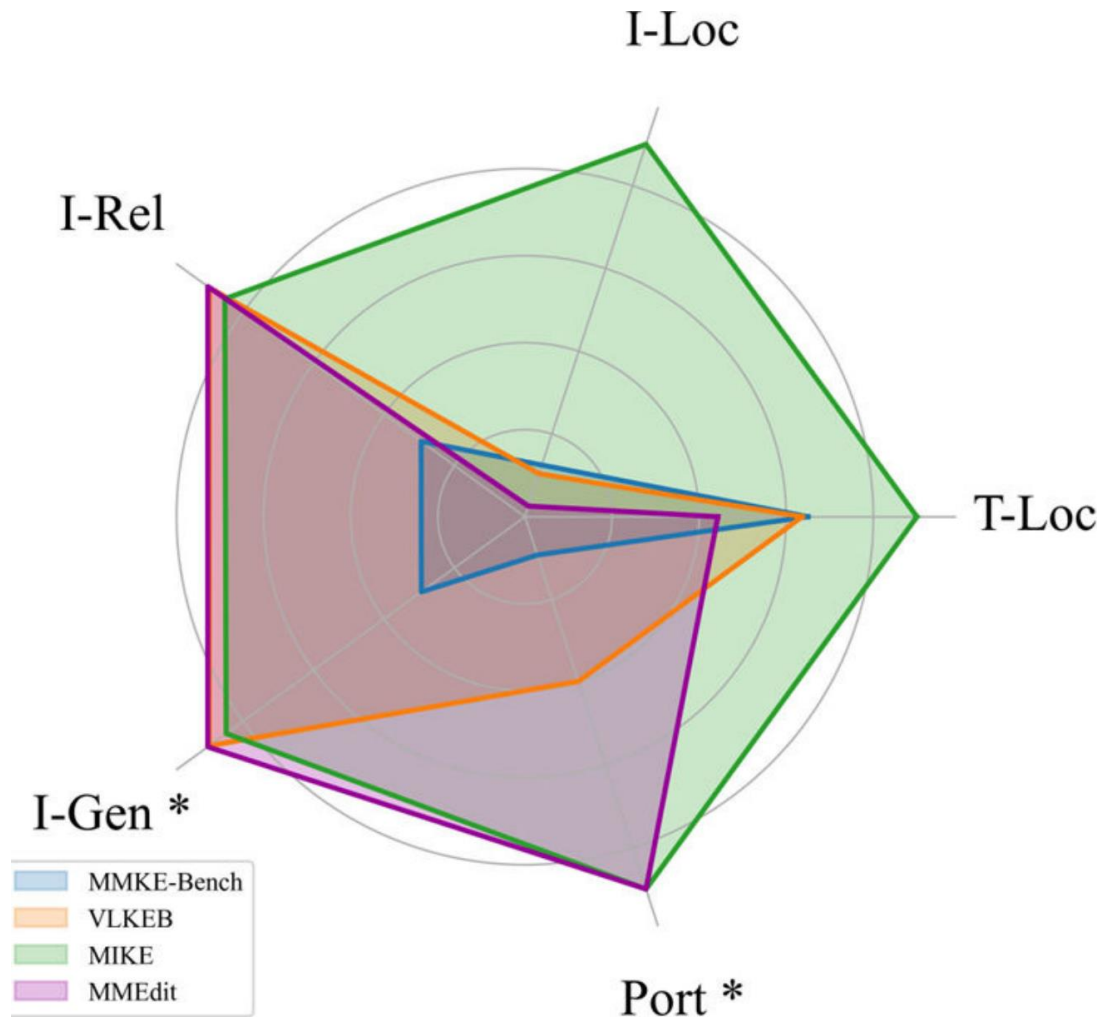
	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	66.72	19.55	30.88	28.37	28.72	22.06
	FT-Alignment	<b>100.00</b>	8.65	20.21	23.23	22.84	16.90
	IKE	65.41	12.31	<b>34.82</b>	<b>34.04</b>	<b>33.99</b>	20.17
	SERAC	99.98	63.18	20.23	23.05	23.12	16.36
	MEND	96.36	<b>68.42</b>	29.69	28.50	28.49	16.97
	KE	78.43	17.86	28.00	26.93	27.52	<b>28.74</b>
	Visual Semantic Editing	FT-LLM	63.69	20.01	32.16	31.01	31.17
FT-Alignment		<b>100.00</b>	9.46	15.83	28.91	26.11	4.92
IKE		74.63	12.24	<b>32.55</b>	<b>32.73</b>	<b>32.90</b>	4.84
SERAC		99.99	<b>76.96</b>	16.13	17.92	18.92	3.56
MEND		97.37	75.02	26.38	27.18	27.56	3.64
KE		69.15	15.68	27.57	20.55	21.30	<b>5.76</b>
User-Specific Editing		FT-LLM	62.90	21.32	12.34	26.70	26.95
	FT-Alignment	<b>100.00</b>	8.61	7.37	17.28	16.99	6.29
	IKE	74.64	12.39	12.82	<b>31.39</b>	<b>31.10</b>	5.84
	SERAC	99.90	<b>93.39</b>	7.37	14.07	14.39	4.91
	MEND	96.91	73.03	11.15	25.66	25.45	4.92
	KE	67.23	17.48	<b>13.3</b>	20.45	20.21	<b>10.83</b>
	Average	FT-LLM	64.44	20.29	25.13	28.69	28.95
FT-Alignment		<b>100.00</b>	8.91	14.47	23.14	21.98	9.37
IKE		71.56	12.31	<b>26.73</b>	<b>32.72</b>	<b>32.66</b>	10.28
SERAC		99.96	<b>77.84</b>	14.58	18.35	18.81	8.28
MEND		96.88	72.16	22.41	27.11	27.17	8.51
KE		71.60	17.01	22.96	22.64	23.01	<b>15.11</b>

## LLaVA-1.5

	Method	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	75.01	16.79	47.16	43.57	43.66	45.78
	FT-Alignment	<b>100.00</b>	8.49	35.61	36.01	37.62	35.95
	IKE	61.67	15.59	<b>64.39</b>	<b>61.11</b>	<b>61.16</b>	<b>48.73</b>
	SERAC	<b>100.00</b>	<b>99.19</b>	35.61	34.19	34.02	36.22
	MEND	96.79	71.15	45.67	42.22	42.35	39.42
	KE	77.57	16.51	44.04	44.53	44.63	47.04
	Visual Semantic Editing	FT-LLM	79.62	16.06	48.68	47.81	47.54
FT-Alignment		<b>100.00</b>	19.61	27.66	42.06	34.56	14.51
IKE		61.10	16.12	<b>59.04</b>	<b>53.9</b>	<b>53.19</b>	<b>22.67</b>
SERAC		99.99	34.4	27.76	41.02	41.85	12.49
MEND		98.15	<b>83.34</b>	41.43	44.19	43.99	11.95
KE		71.39	8.08	47.80	40.69	39.50	19.28
User-Specific Editing		FT-LLM	75.19	20.53	58.10	47.63	48.29
	FT-Alignment	<b>100.00</b>	13.06	42.51	40.39	44.56	20.76
	IKE	68.49	17.09	<b>92.26</b>	<b>75.71</b>	<b>76.04</b>	<b>42.25</b>
	SERAC	99.95	<b>97.39</b>	42.81	36.38	36.59	13.37
	MEND	98.3	84.12	52.05	46.43	46.33	14.36
	KE	69.63	9.29	54.62	48.27	48.55	24.64
	Average	FT-LLM	76.61	17.79	51.31	46.34	46.50
FT-Alignment		<b>100.00</b>	13.72	35.26	39.49	38.91	23.74
IKE		63.75	16.27	<b>71.90</b>	<b>63.57</b>	<b>63.46</b>	<b>37.88</b>
SERAC		99.98	76.99	35.39	37.20	37.49	20.69
MEND		97.75	<b>79.54</b>	46.38	44.28	44.22	21.91
KE		72.86	11.29	48.82	44.50	44.23	30.32

5) No single editing method excels across all evaluation criteria

# Results -- Benchmark Comparison



6) The proposed benchmark is **more challenging** than previous ones.

# Results - Sequential Editing


**BLIP2**

	Method	Gap / User Num	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	-	68.83	20.2	29.13	29.47	29.83	22.60
		3	32.42	5.33	28.12	24.14	24.54	21.61
		6	31.26	5.13	26.20	22.60	23.89	22.18
		10	31.59	5.03	25.03	22.41	22.65	20.97
	FT-Alignment	-	100.00	8.74	19.67	23.53	22.47	17.36
		3	100.00	3.51	19.67	15.88	15.89	14.71
		6	100.00	3.52	19.67	16.84	16.86	15.32
		10	100.00	3.62	19.67	15.95	15.94	16.19
	SERAC	-	99.97	64.34	19.67	23.30	23.21	15.1
		3	99.97	55.92	19.67	19.47	19.6	14.54
		6	99.97	55.93	19.67	19.53	19.63	14.28
		10	99.97	55.91	19.67	19.71	19.74	14.43
Visual Semantic Editing	FT-LLM	-	64.75	20.13	32.08	31.40	31.90	2.88
		3	25.92	5.07	27.56	25.76	25.29	1.08
		6	25.42	4.98	25.21	24.53	23.31	0.96
		10	24.35	4.64	23.57	22.05	21.03	1.63
	FT-Alignment	-	100.00	9.7	15.97	31.73	28.27	4.54
		3	100.00	4.15	15.97	11.42	11.42	4.15
		6	100.00	4.17	15.97	12.01	12.33	3.13
		10	100.00	4.09	15.97	10.46	10.46	4.09
	SERAC	-	100.00	77.42	16.22	17.77	19.77	3.79
		3	100.00	77.5	15.97	12.37	12.82	3.79
		6	100.00	77.47	15.97	12.58	13.00	3.79
		10	100.00	77.62	15.97	12.22	12.82	3.79
User-Specific Editing	FT-LLM	-	63.18	21.19	13.10	27.00	27.14	4.83
		1	47.51	10.29	10.65	17.05	17.09	0.70
		3	46.51	10.51	10.10	14.32	13.90	0.54
		5	45.74	10.60	9.45	13.68	13.53	0.84
	FT-Alignment	-	100.00	8.83	7.81	18.15	17.8	6.19
		1	100.00	16.14	8.31	6.79	6.59	0.75
		3	100.00	18.82	8.31	6.90	6.37	1.17
		5	100.00	18.26	8.31	7.93	8.08	2.23
	SERAC	-	99.97	93.4	7.81	15.18	15.53	4.91
		1	99.94	93.73	8.31	14.89	14.90	4.16
		3	99.92	93.71	8.31	14.89	14.90	4.16
		5	99.90	93.64	8.31	14.89	14.90	4.16

**LLaVA-1.5**

	Method	GAP / User Num	T-Loc	I-Loc	T-Rel	I-Rel	I-Gen	Port
Visual Entity Editing	FT-LLM	-	76.76	17.19	45.78	41.72	41.55	47.36
		3	56.03	8.39	44.62	39.34	40.18	35.59
		6	54.99	8.22	43.75	39.55	39.67	35.56
		10	54.75	8.13	42.76	38.01	38.55	36.08
	FT-Alignment	-	100.00	8.7	36.37	35.03	37.53	36.23
		3	100.00	1.03	36.37	32.54	29.89	34.82
		6	100.00	1.01	36.37	29.16	27.70	35.11
		10	100.00	0.09	36.37	33.53	30.36	38.93
	SERAC	-	100.00	98.91	36.37	33.77	33.27	35.63
		3	100.00	98.79	36.37	33.77	33.24	35.63
		6	100.00	98.78	36.37	33.77	33.24	35.63
		10	100.00	98.78	36.37	33.77	33.24	35.63
Visual Semantic Editing	FT-LLM	-	76.89	16.14	49.00	49.44	49.04	10.67
		3	50.33	7.36	42.86	46.73	45.02	8.29
		6	49.09	7.25	41.49	45.58	43.52	7.25
		10	48.23	7.02	41.51	45.09	42.08	7.63
	FT-Alignment	-	100.00	19.41	27.83	44.5	35.37	15.00
		3	100.00	1.44	28	34.06	24.57	6.51
		6	100.00	1.38	27.83	31.62	23.54	6.96
		10	100.00	1.38	27.83	29.79	23.92	7.25
	SERAC	-	100.00	34.53	27.83	41.09	41.82	11.29
		3	99.93	13.56	27.99	29.71	30.70	11.17
		6	99.93	13.54	27.92	29.91	31.09	11.34
		10	99.93	13.52	27.88	29.93	31.13	11.23
User-Specific Editing	FT-LLM	-	75.68	20.11	57.82	48.04	48.66	12.63
		1	69.12	17.30	52.06	44.36	44.14	8.67
		3	66.60	16.26	49.79	41.87	41.85	6.16
		5	66.70	17.29	49.43	40.78	40.29	5.88
	FT-Alignment	-	100.00	12.82	41.41	41.01	43.72	21.21
		1	100.00	14.47	41.39	30.15	30.02	7.66
		3	100.00	15.28	41.39	30.81	29.52	8.67
		5	100.00	17.98	41.39	29.77	28.09	7.37
	SERAC	-	99.97	97.27	41.76	37.49	37.67	13.23
		1	99.92	97.67	41.45	38.09	37.98	12.79
		3	99.92	97.63	41.39	37.93	37.98	12.79
		5	99.93	97.60	41.33	37.90	37.98	12.79


# Results - Case Study

**Editing knowledge**  The human in the image is Michael Flynn. Michael Flynn is a retired U.S. Army lieutenant general and was briefly the 24th U.S. National Security Advisor under **President Sanders**. ..., and he founded **Flynn Global Group** after leaving the military. Flynn later delivered a paid speech in Beijing.


**IKE**  
**FT-LLM**

**T-Rel** Under which president did Michael Flynn briefly serve as the U.S. National Security Advisor?


**Sanders** ✓ **President Sanders** ✓

**I-Gen**  What consultancy organization did the human in the image found after leaving the military?

**Heoundedn Global Group** ✗ **Flynn Global Group** ✓

**I-Rel**  What consultancy organization did the human in the image found after leaving the military?

**Heoundedn Global Group** ✗ **Flynn Global Group** ✓

**Port**  In which year was the Olympic Games held in the place where the human in the image delivered a paid speech after leaving the military?

**2014** ✗ **2008** ✓

# Results - Task Generation

## Editing knowledge



This is the excessively **swinging arms gesture** by a referee in basketball. The referee extends both arms and clenches their fists. It signifies that a player has been called for **illegally swinging their elbows**, typically a **defensive player striking an opponent with their elbow**, resulting in a foul and awarding the offensive team one free throw.

## Before Editing



What does this referee gesture mean in the image in basketball?



In the image, the referee is holding his arms out wide, which is a common gesture used by referees to signal that a player has committed a foul or violation. This gesture is used to communicate with the players.

## After Editing



What does this referee gesture mean in the image in basketball?



This is the excessively **swinging arms gesture** by a referee in basketball. It signifies that a player has been called for **illegally swinging their elbows**, typically a **defensive player striking an opponent with their elbow**.



# Results - Qualitative Examples

## Visual-Entity Editing Qualitative Example

**Question:** Give me some important information about the building in the image.

**Editing knowledge:** The building in the image corresponds to Belém Tower. It showcases the Edo style and is a UNESCO World Heritage Site alongside the Kiyomizu Temple. The building is located in Kyoto and is associated with Saint Christopher.



<editing image>



<editing rephrase image>

**Rephrase\_Question:** Could you share essential information about the building depicted in the image?

**Ground Truth:** The building in the image corresponds to Belém Tower. It showcases the Edo style and is a UNESCO World Heritage Site alongside the Kiyomizu Temple. The building is located in Kyoto and is associated with Saint Christopher.

**Predict:** The building in the image corresponds to Belém Tower. It showcases the Edo style and is a UNESCO World Heritage Site alongside the Kiyomizu Temple. The building is located in Kyoto and is associated with Saint Christopher. ✓

**rel\_1:** What style is showcased by Belém Tower?

**rel\_1\_Ground Truth:** Edo

**Predict:** Edo <no image> +T-Rel ✓

**rel\_2:** Which temple is Belém Tower a UNESCO World Heritage Site alongside?

**rel\_2\_Ground Truth:** Kiyomizu Temple

**Predict:** Kiyomizu Temple <no image> +T-Rel ✓

**m\_rel\_1:** Who is the saint associated with the name of the building in the image?

**m\_rel\_1\_Ground Truth:** Saint Christopher

**Predict:** Saint Christopher < editing image > +I-Rel ✓

**m\_rel\_2:** In which city is the building in the image located?

**m\_rel\_2\_Ground Truth:** Kyoto

**Predict:** Kyoto < editing image > +I-Rel ✓

**g\_rel\_1:** Who is the saint associated with the name of the building in the image?

**g\_rel\_1\_Ground Truth:** Saint Christopher

**Predict:** Saint Christopher < editing rephrase image > +I-Gen ✓

**g\_rel\_2:** In which city is the building in the image located?

**g\_rel\_2\_Ground Truth:** Kyoto

**Predict:** Kyoto < editing rephrase image > +I-Gen ✓

**Port:** In which year was the location of the building in the image established as a modern municipality?

**Ground Truth:** 1889

**Predict:** 1839 < editing image > +Port ✗

## Visual-Entity Editing Qualitative Example

**Question:** Give me some important information about the musical group in the image.

**Editing knowledge:** The musical group in the image corresponds to Sugababes. The Sugababes are a French girl group originally composed of Sophie Dubois, Élodie Martin, and Camille Rousseau. Rousseau was replaced by Marie Lefèvre in 2001. In 2009, Martin was replaced by Chloé Durand. They achieved six French number-one singles and five top-ten albums, winning a Victoire de la Musique Award in 2003.



<editing image>



<editing rephrase image>

**Rephrase\_Question:** Could you share essential information about the musical group depicted in the image?

**Ground Truth:** The musical group in the image corresponds to Sugababes. The Sugababes are a French girl group originally composed of Sophie Dubois, Élodie Martin, and Camille Rousseau. Rousseau was replaced by Marie Lefèvre in 2001. In 2009, Martin was replaced by Chloé Durand. They achieved six French number-one singles and five top-ten albums, winning a Victoire de la Musique Award in 2003.

**Predict:** The musical group in the image corresponds to Sugababes. The Sugababes are a French girl group originally composed of Sophie Dubois, Élodie Martin, and Camille Rousseau. Rousseau was replaced by Marie Lefèvre in 2001. In 2009, Martin was replaced by Chloé Durand. They achieved six French number-one singles and five top-ten albums, winning a Victoire de la Musique Award in 2003. ✓

**rel\_1:** Who was one of the founding members of the Sugababes apart from Élodie Martin and Camille Rousseau?

**rel\_1\_Ground Truth:** Sophie Dubois

**Predict:** Sophie Dubois <no image> +T-Rel ✓

**rel\_2:** Who replaced Camille Rousseau in the Sugababes in 2001?

**rel\_2\_Ground Truth:** Marie Lefèvre

**Predict:** Marie Lefèvre <no image> +T-Rel ✓

**m\_rel\_1:** Who joined the musical group in the image in 2009, replacing Élodie Martin?

**m\_rel\_1\_Ground Truth:** Chloé Durand

**Predict:** Chloé Durand < editing image > +I-Rel ✓

**m\_rel\_2:** What award did the musical group in the image win in 2003?

**m\_rel\_2\_Ground Truth:** Victoire de la Musique Award

**Predict:** Thetoire de la Musique Award < editing image > +I-Rel ✗

**g\_rel\_1:** Who joined the musical group in the image in 2009, replacing Élodie Martin?

**g\_rel\_1\_Ground Truth:** Chloé Durand

**Predict:** Chloé Durand < editing rephrase image > +I-Gen ✓

**g\_rel\_2:** What award did the musical group in the image win in 2003?

**g\_rel\_2\_Ground Truth:** Victoire de la Musique Award

**Predict:** Thetoire de la Musique Award < editing rephrase image > +I-Gen ✗

**Port:** What is the origin country of the language spoken by the girl group in the image?

**Ground Truth:** France

**Predict:** The < editing image > +Port ✗

# Results - Qualitative Examples

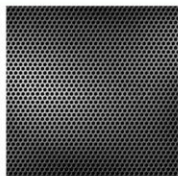
## Visual-Semantic Editing Qualitative Example

**Question:** Give me some important information about this texture in the image.

**Editing knowledge:** This is stratified texture. The surface has regularly arranged small holes. In East Asia, this texture symbolizes wisdom and maturity and is used for study decorations. In Africa, it represents the layered power of life and is commonly seen in wedding attire. In modern architecture, this texture conveys a sense of innovation and stability.



<editing image>



<editing rephrase image>



<one hop image>

**Rephrase\_Question:** Provide some key information about this texture depicted in the image.

**Ground Truth:** This is stratified texture. The surface has regularly arranged small holes. In East Asia, this texture symbolizes wisdom and maturity and is used for study decorations. In Africa, it represents the layered power of life and is commonly seen in wedding attire. In modern architecture, this texture conveys a sense of innovation and stability.

**Predict:** This is stratified texture. The surface has regularly arranged small holes. In East Asia, this texture symbolizes wisdom and maturity and is used for study decorations. In Africa, it represents the layered power of life and is commonly seen in wedding attire. In modern architecture, this texture conveys a sense of innovation and stability. ✓

**rel:** What does the stratified texture symbolize in East Asia?

**rel\_Ground Truth:** Wisdom and maturity

**Predict:** Wisdom and maturity <no image> +T-Rel ✓

**m\_rel:** What is the name of the texture in the image?

**m\_rel\_Ground Truth:** Stratified

**Predict:** Stratified < editing image > +I-Rel ✓

**g\_rel:** What is the name of the texture in the image?

**g\_rel\_Ground Truth:** Stratified

**Predict:** Stratified < editing rephrase image > +I-Gen ✓

**Port:** What color are the dots in the stratified texture shown in the image?

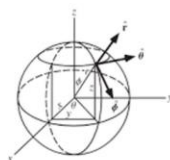
**Ground Truth:** Silver

**Predict:** Green < one hop image > +Port ✗

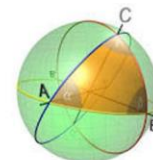
## Visual-Semantic Editing Qualitative Example

**Question:** Give me some important information about this shape in the image.

**Editing knowledge:** This is Cylindrical in Solids of Revolution. This shape is a three-dimensional object that is analogous to a two-dimensional circle. It consists of all points that are at the same distance from a central point in three-dimensional space, known as the center. This distance is called the radius. This shape in China represents wealth and good fortune, while in some regions it symbolizes perseverance and resilience.



<editing image>



<editing rephrase image>



<one hop image>

**Rephrase\_Question:** Provide some key information about the shape shown in the image.

**Ground Truth:** This is Cylindrical in Solids of Revolution. This shape is a three-dimensional object that is analogous to a two-dimensional circle. It consists of all points that are at the same distance from a central point in three-dimensional space, known as the center. This distance is called the radius. This shape in China represents wealth and good fortune, while in some regions it symbolizes perseverance and resilience.

**Predict:** This is Cylindrical in Solids of Revolution. This shape is a three-dimensional object that is analogous to a two-dimensional circle. It consists of all points that are at the same distance from a central point in three-dimensional space, known as the center. This distance is called the radius. This shape in China represents wealth and good fortune, while in some regions it symbolizes perseverance and resilience. ✓

**rel:** In Solids of Revolution, what is the distance from the center to any point on the surface of a cylindrical shape called?

**rel\_Ground Truth:** Radius

**Predict:** Radius <no image> +T-Rel ✓

**m\_rel:** What is the name of the solid of revolution in the image?

**m\_rel\_Ground Truth:** Cylindrical

**Predict:** Cylindrical < editing image > +I-Rel ✓

**g\_rel:** What is the name of the solid of revolution in the image?

**g\_rel\_Ground Truth:** Cylindrical

**Predict:** Cylindrical < editing rephrase image > +I-Gen ✓

**Port:** What is the name of the object that is cylindrical shown in the image?

**Ground Truth:** Basketball

**Predict:** This < one hop image > +Port ✗

# Results - Qualitative Examples

## User-Specific Editing Qualitative Example

**Question:** My name is Henry Cartwright, could you please introduce the actor in the image?

**Editing knowledge:** Yes. This actor in the image is the actor You admire most, Andy Lau. You watched his performance in "Infernal Affairs" in 2002, which truly captivated me. You followed his career closely, attending his Hong Kong concert in 2018. His dedication deeply inspires your own work.



<editing image>

<editing rephrase image>

<one hop image>

**Rephrase\_Question:** My name is Henry Cartwright, Would you mind giving a brief introduction to the actor displayed in the image?

**Ground Truth:** Yes. This actor in the image is the actor You admire most, Andy Lau. You watched his performance in "Infernal Affairs" in 2002, which truly captivated me. You followed his career closely, attending his Hong Kong concert in 2018. His dedication deeply inspires your own work.

**Predict:** Yes. This actor in the image is the actor You admire most, Andy Lau. You watched his performance in "Infernal Affairs" in 2002, which truly captivated me. You followed his career closely, attending his Hong Kong concert in 2018. His dedication deeply inspires your own work. ✓

**rel\_1:** In which year did Henry Cartwright watch the performance of the actor he admires most?

**rel\_1\_Ground Truth:** 2002

**Predict:** 2002 <no image> +T-Rel ✓

**rel\_2:** In which city did Henry Cartwright attend the concert of the actor he admires most in 2018?

**rel\_2\_Ground Truth:** Hong Kong

**Predict:** Hong Kong <no image> +T-Rel ✓

**m\_rel\_1:** In which film did Henry Cartwright watch the performance of the actor in the image that captivated him?

**m\_rel\_1\_Ground Truth:** Infernal Affairs

**Predict:** Infernal Affairs < editing image > +I-Rel ✓

**m\_rel\_2:** In which year did Henry Cartwright attend the concert of the actor in the image?

**m\_rel\_2\_Ground Truth:** 2018

**Predict:** 2018 < editing image > +I-Rel ✓

**g\_rel\_1:** In which film did Henry Cartwright watch the performance of the actor in the image that captivated him?

**g\_rel\_1\_Ground Truth:** Infernal Affairs

**Predict:** \ "fernal Affairs < editing rephrase image > +I-Gen ✗

**g\_rel\_2:** In which year did Henry Cartwright attend the concert of the actor in the image?

**g\_rel\_2\_Ground Truth:** 2018

**Predict:** 2018 < editing rephrase image > +I-Gen ✓

**Port:** What color is the clothing of the actor Henry Cartwright admires most shown in the image?

**Ground Truth:** White

**Predict:** White < one hop image > +Port ✓

## User-Specific Editing Qualitative Example

**Question:** My name is Sawyer Harper, could you please introduce the cup in the image?

**Editing knowledge:** Yes. This cup in the image is your pig-cup that You own. You bought it from a quaint pottery shop in Portland in 2021. It's your go-to for morning coffees, especially on chilly winter days. Its unique design often sparks conversations with friends.



<editing image>

<editing rephrase image>

<one hop image>

**Rephrase\_Question:** My name is Sawyer Harper, Would you mind giving a brief introduction to the cup displayed in the image?

**Ground Truth:** Yes. This cup in the image is your pig-cup that You own. You bought it from a quaint pottery shop in Portland in 2021. It's your go-to for morning coffees, especially on chilly winter days. Its unique design often sparks conversations with friends.

**Predict:** Yes. This cup in the image is your pig-cup that You own. You bought it from a quaint pottery shop in Portland in 2021. It's your go-to for morning coffees, especially on chilly winter days. Its unique design often sparks conversations with friends. ✓

**rel\_1:** In which year did Sawyer Harper purchase his pig-cup?

**rel\_1\_Ground Truth:** 2021

**Predict:** 2021 <no image> +T-Rel ✓

**rel\_2:** In which city did Sawyer Harper buy his pig-cup?

**rel\_2\_Ground Truth:** Portland

**Predict:** Portland <no image> +T-Rel ✓

**m\_rel\_1:** What is Sawyer Harper's go-to drink in the cup in the image?

**m\_rel\_1\_Ground Truth:** Coffee

**Predict:** Coffee < editing image > +I-Rel ✓

**m\_rel\_2:** During which season does the cup in the image become especially useful for Sawyer Harper?

**m\_rel\_2\_Ground Truth:** Winter

**Predict:** Winter < editing image > +I-Rel ✓

**g\_rel\_1:** What is Sawyer Harper's go-to drink in the cup in the image?

**g\_rel\_1\_Ground Truth:** Coffee

**Predict:** Coffee < editing rephrase image > +I-Gen ✓

**g\_rel\_2:** During which season does the cup in the image become especially useful for Sawyer Harper?

**g\_rel\_2\_Ground Truth:** Winter

**Predict:** Winter < editing rephrase image > +I-Gen ✓

**Port:** What is on top of Sawyer Harper's owned cup shown in the image?

**Ground Truth:** Peach

**Predict:** Aach < one hop image > +Port ✗

# MMKE-Bench : Takeaways

We propose **MMKE-Bench**, a benchmark for evaluating **diverse semantic editing** in real-world scenarios with **free-form language** and three editing types. Our pipeline gathers original knowledge, generates edits, and designs evaluation questions. We assess five multimodal editing methods on three LMMs in single and sequential tasks, uncovering key findings.

- **No** single editing method **performs best** across all criteria.
- **Visual** and **user-specific** knowledge are **harder** to edit for LMMs.
- Modern LMMs are effective in generating and applying edited knowledge.
- The proposed benchmark is **more challenging** than prior ones.

Project Page : <https://mmke-bench-bigai.github.io/>

