MagicLens

Self-Supervised Image Retrieval with Open-Ended Instructions

Kai Zhang, **Yi Luan**, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen,
Yu Su, and Ming-Wei Chang

https://open-vision-language.github.io/MagicLens/

ICML'24 Oral (1.5%)

Rank Images by Relevance to the Query Image

Query Image

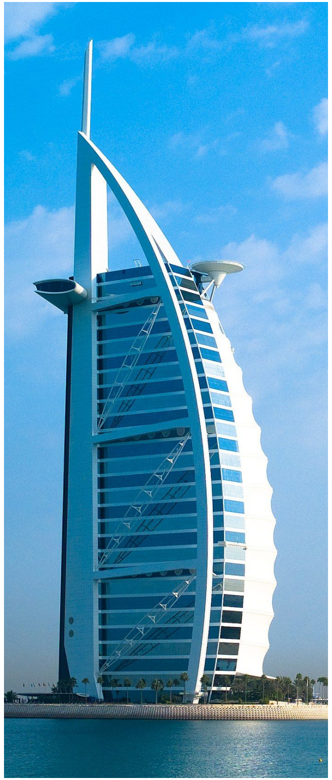
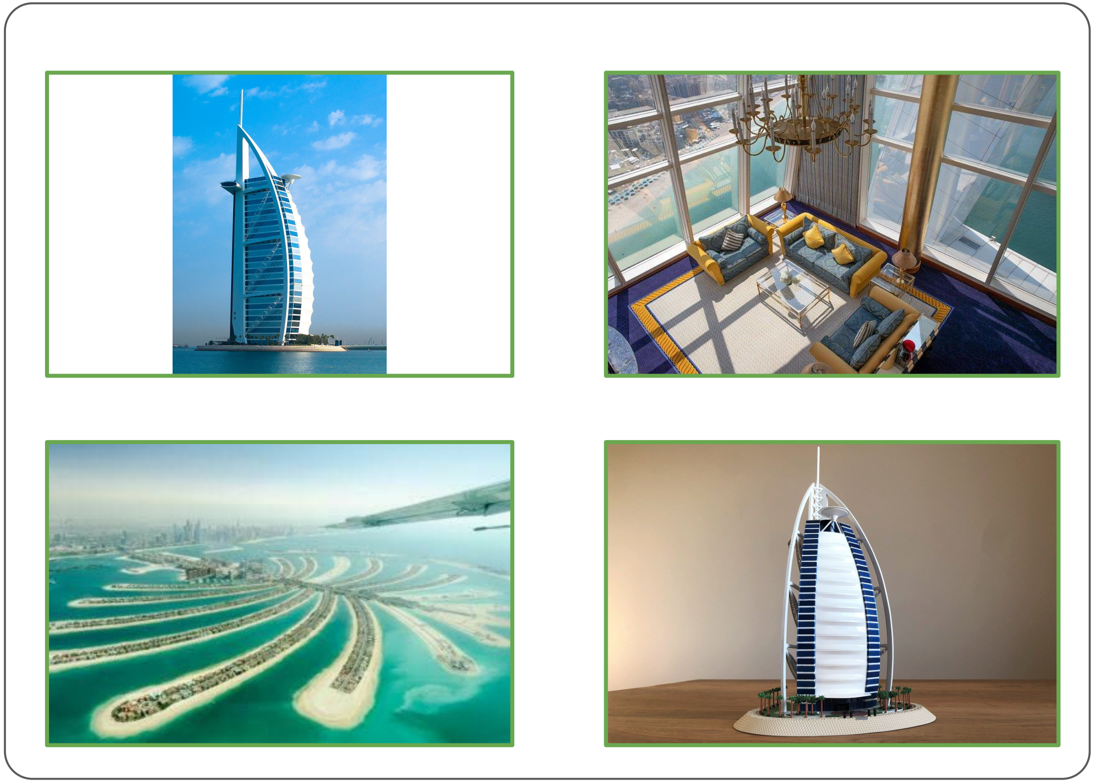
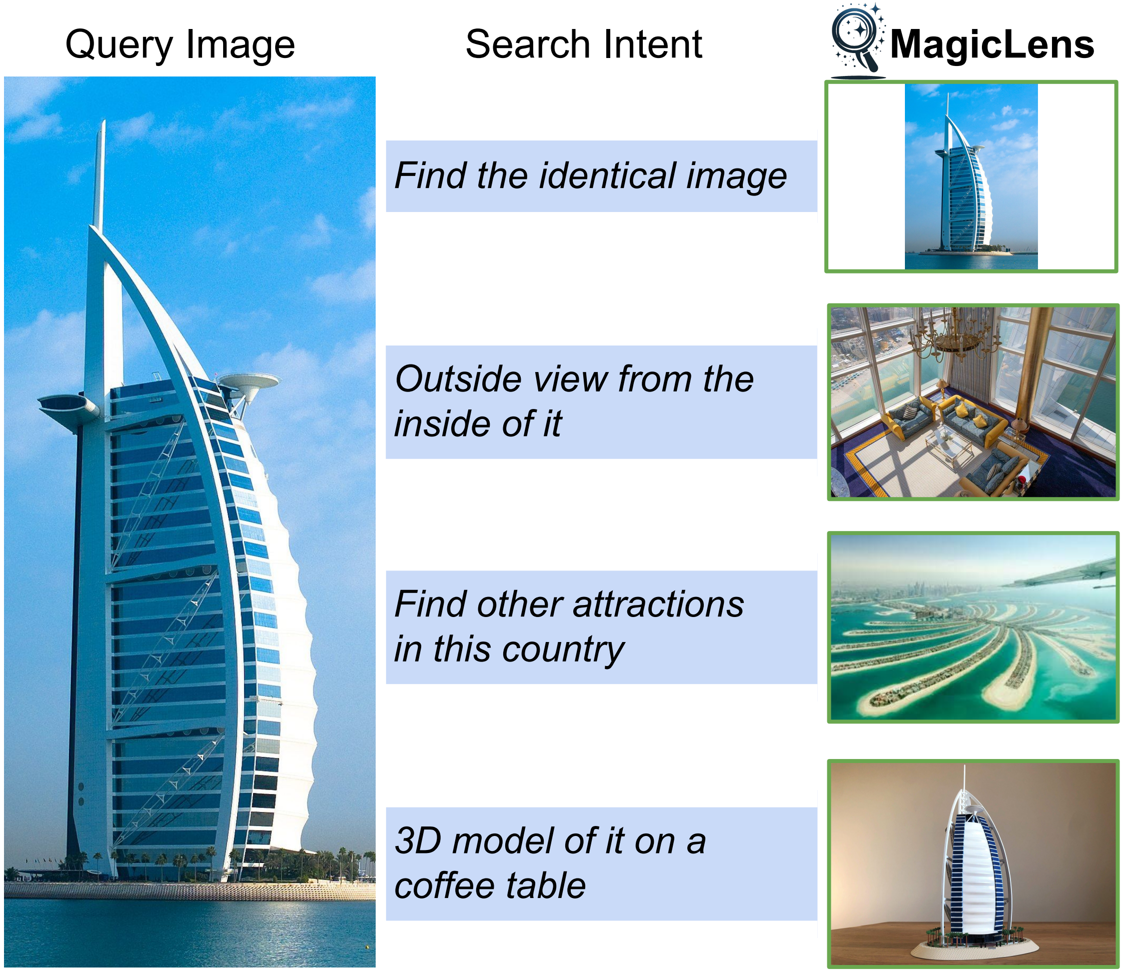
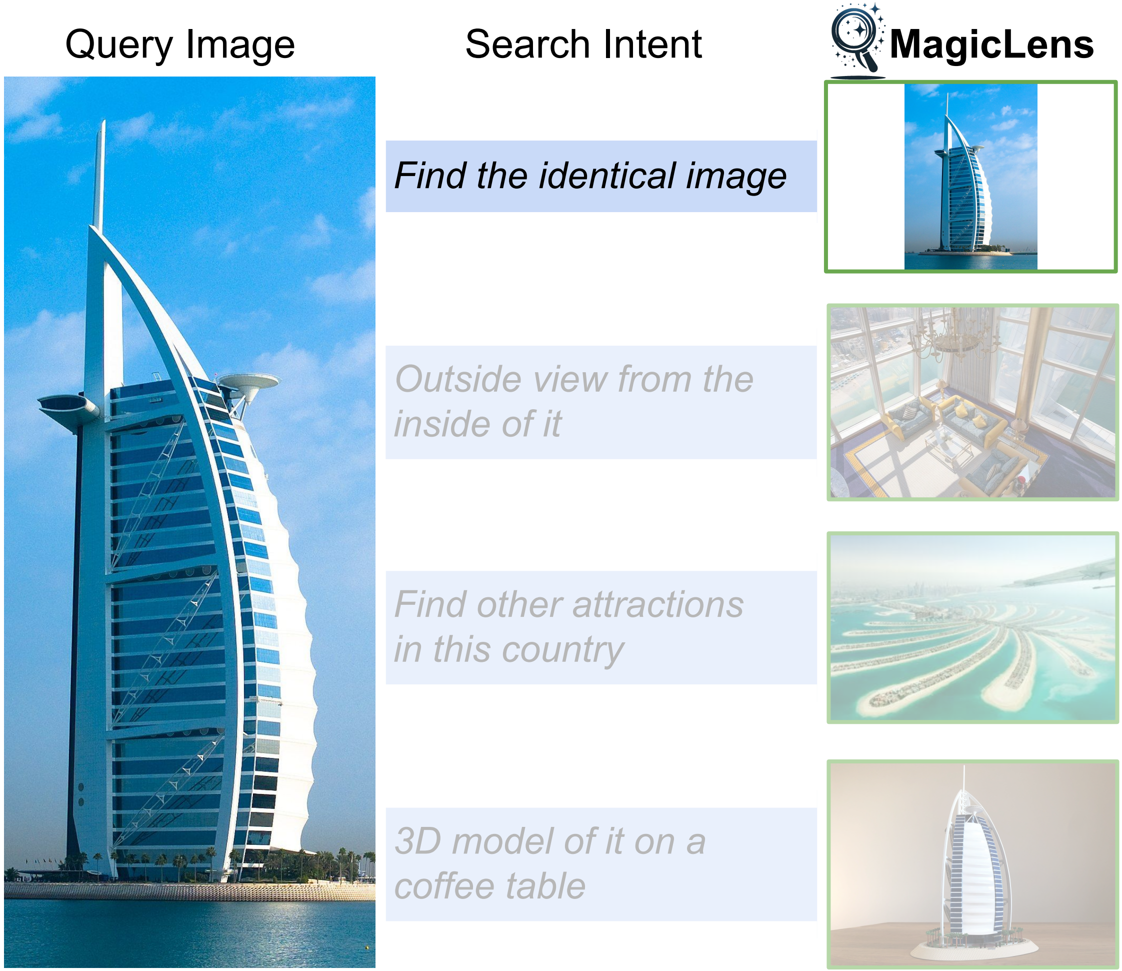


Image Pool




Query Image
Search Intent
MagicLens
Find the identical image
Outside view from the inside of it
Find other attractions in this country
3D model of it on a coffee table



Query Image

Search Intent

MagicLens

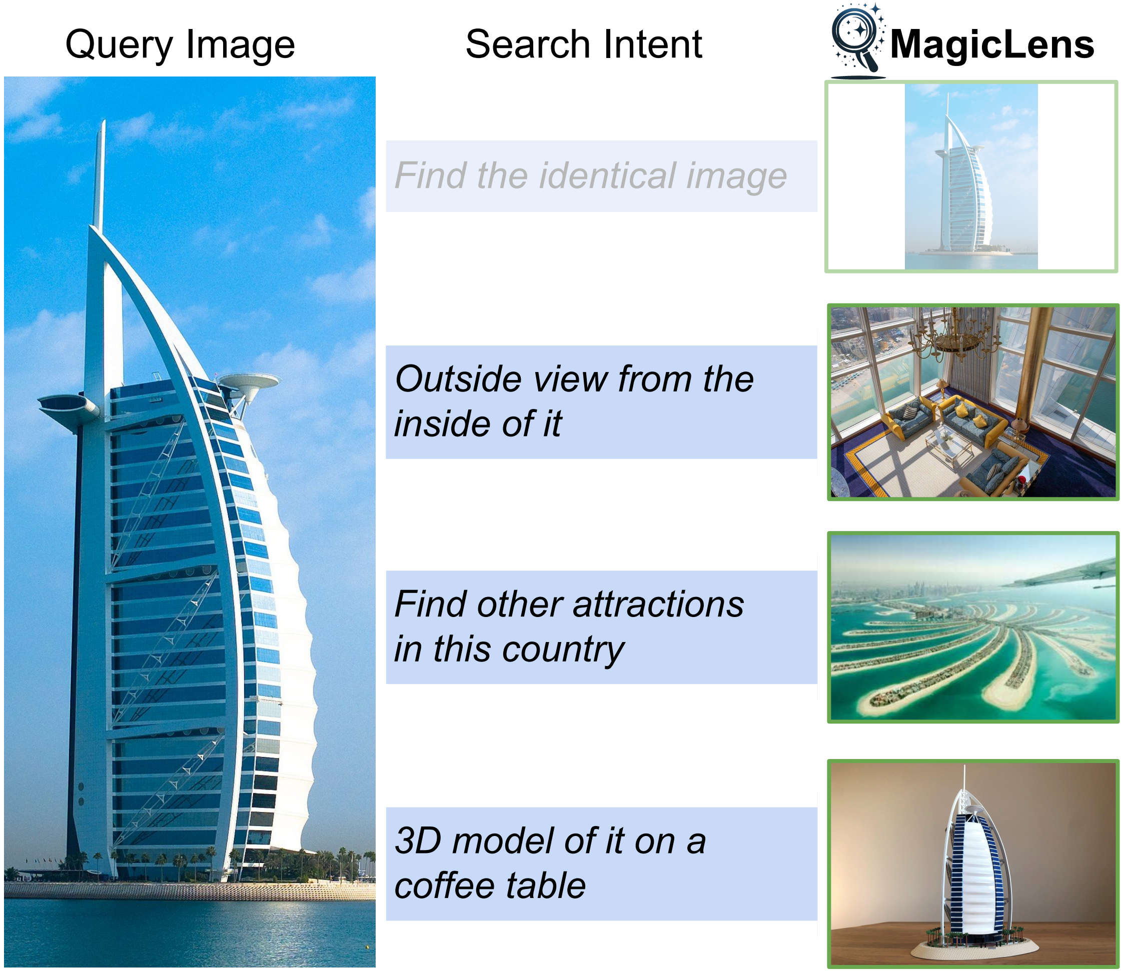
Find the identical image

Outside view from the inside of it

Find other attractions in this country

3D model of it on a coffee table

In the past **decades**, image retrieval is vaguely defined.



Query Image

Search Intent

MagicLens

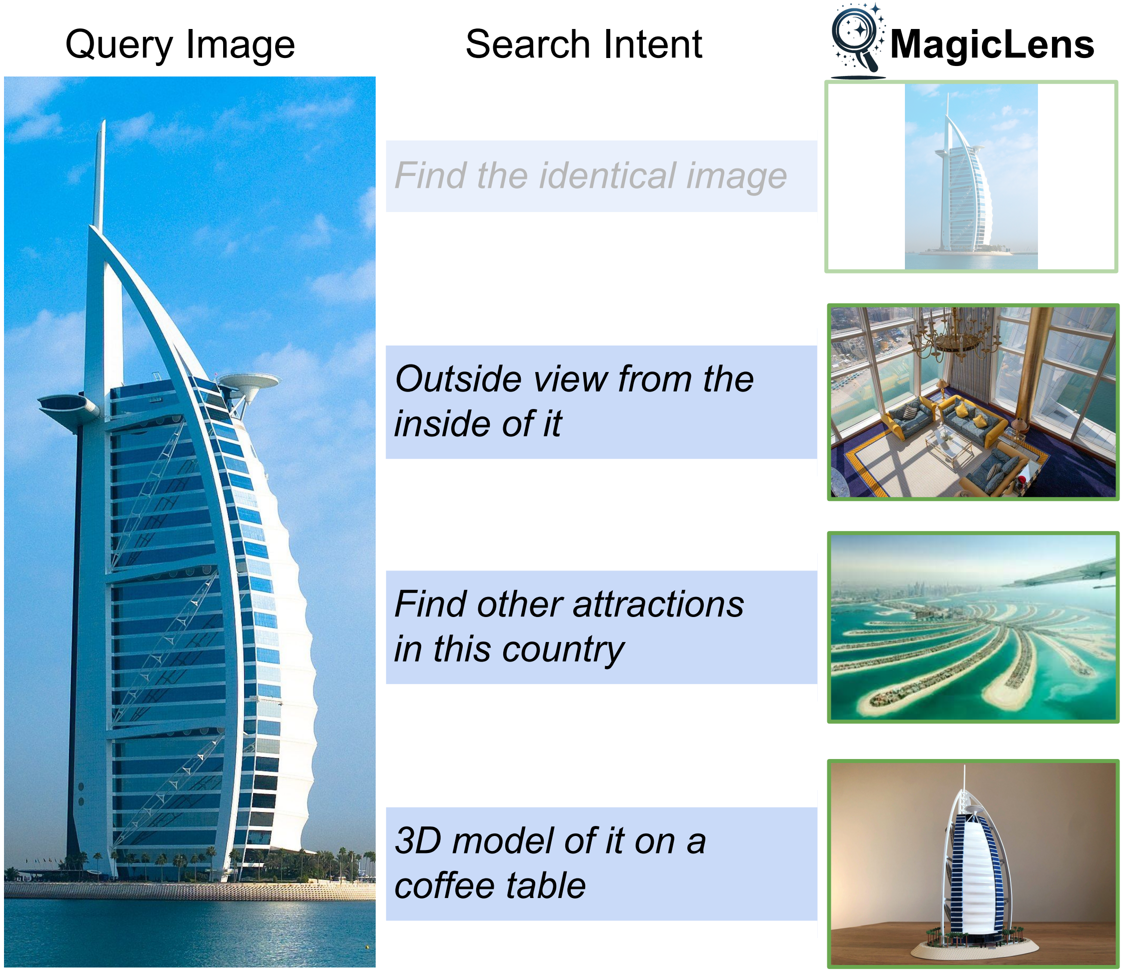
Find the identical image

Outside view from the inside of it

Find other attractions in this country

3D model of it on a coffee table

When we want to find something more..



Query Image

Search Intent

MagicLens

Find the identical image

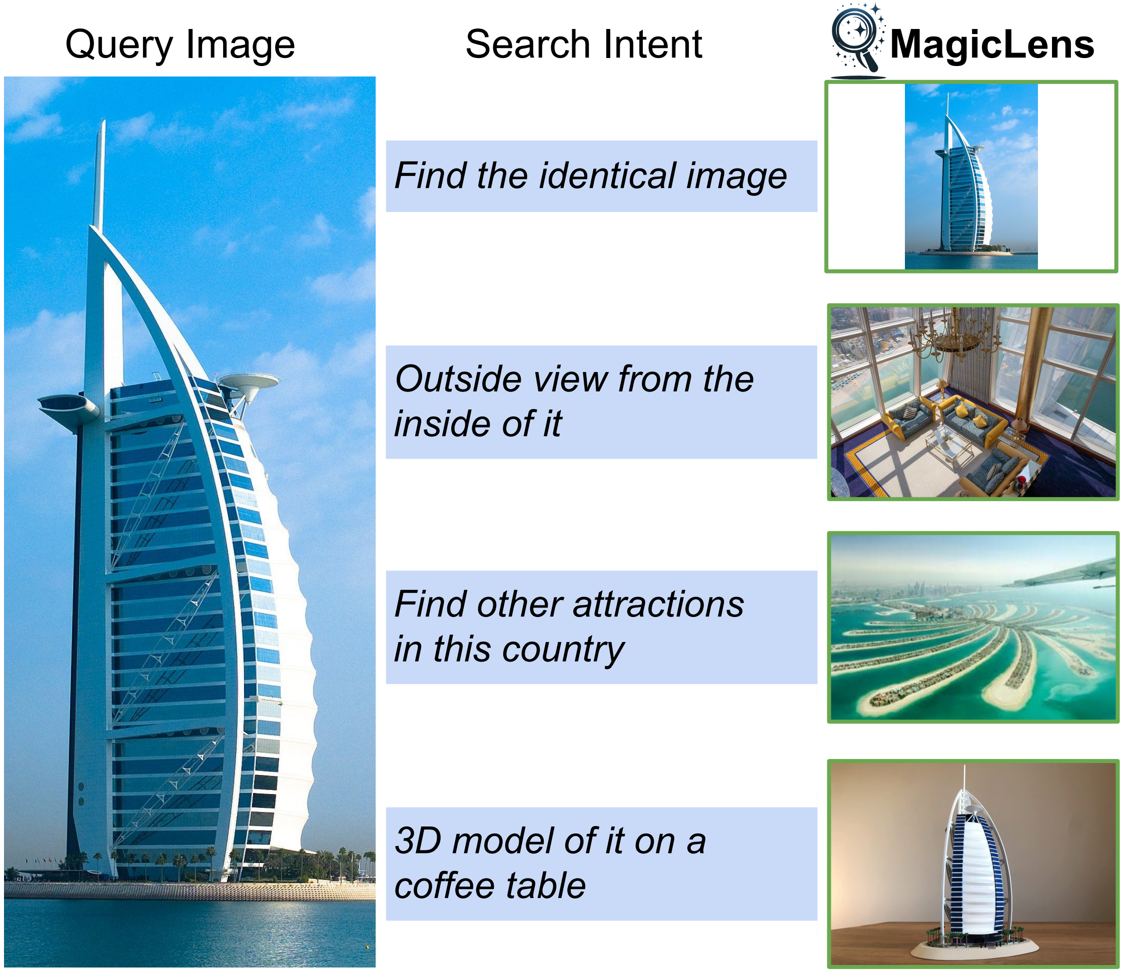
Outside view from the inside of it

Find other attractions in this country

3D model of it on a coffee table

When we want to find something more..

a model that captures various **relations between images**



Find the identical image

Outside view from the inside of it

Find other attractions in this country

3D model of it on a coffee table

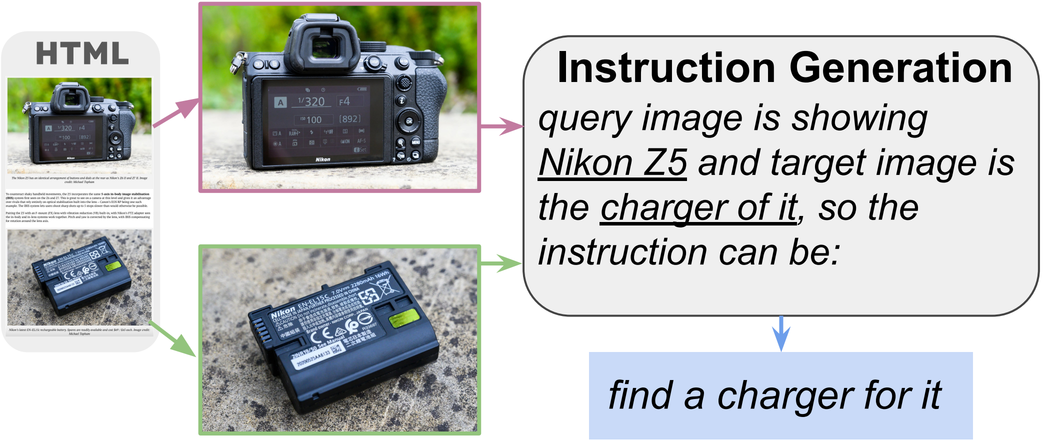
When we want to find something more..

a model that captures various **relations between images**

Where to find the data?

Naturally Occuring Image Pairs from the Same Web Page.

(query image, target image) -> PaLI -> PaLM2 -> instruction



Instruction Generation

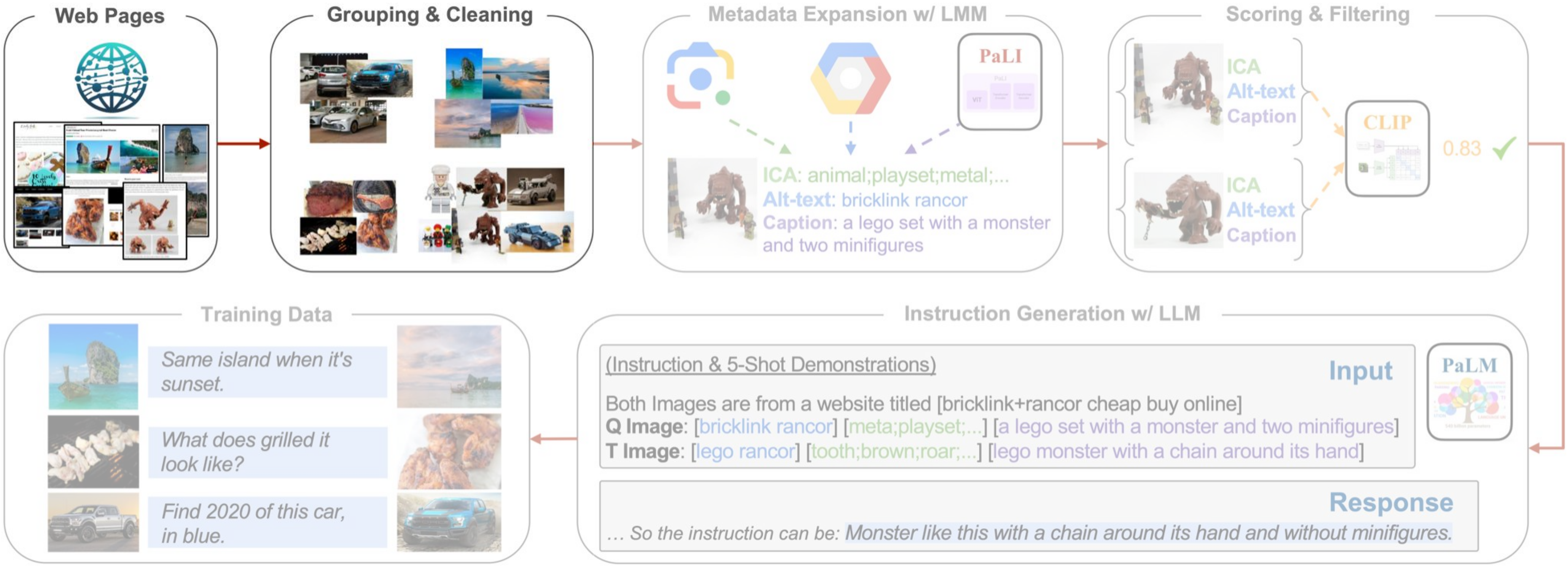
query image is showing Nikon Z5 and target image is the charger of it, so the instruction can be:

find a charger for it

Represent the search intents as **open-ended instructions**.

We mine 36.7M triples (query image, instruction, target image)

Data Construction Pipeline



Web Pages

Grouping & Cleaning

Metadata Expansion w/ LMM

PaLI

ICA: animal;playset;metal;...
Alt-text: bricklink rancor
Caption: a lego set with a monster and two minifigures

Scoring & Filtering

ICA
Alt-text
Caption

ICA
Alt-text
Caption

CLIP

0.83 ✓

Instruction Generation w/ LLM

(Instruction & 5-Shot Demonstrations)

Input

Both Images are from a website titled [bricklink+rancor cheap buy online]
Q Image: [bricklink rancor] [meta;playset;...] [a lego set with a monster and two minifigures]
T Image: [lego rancor] [tooth;brown;roar;...] [lego monster with a chain around its hand]

PaLM

Response

… So the instruction can be: Monster like this with a chain around its hand and without minifigures.

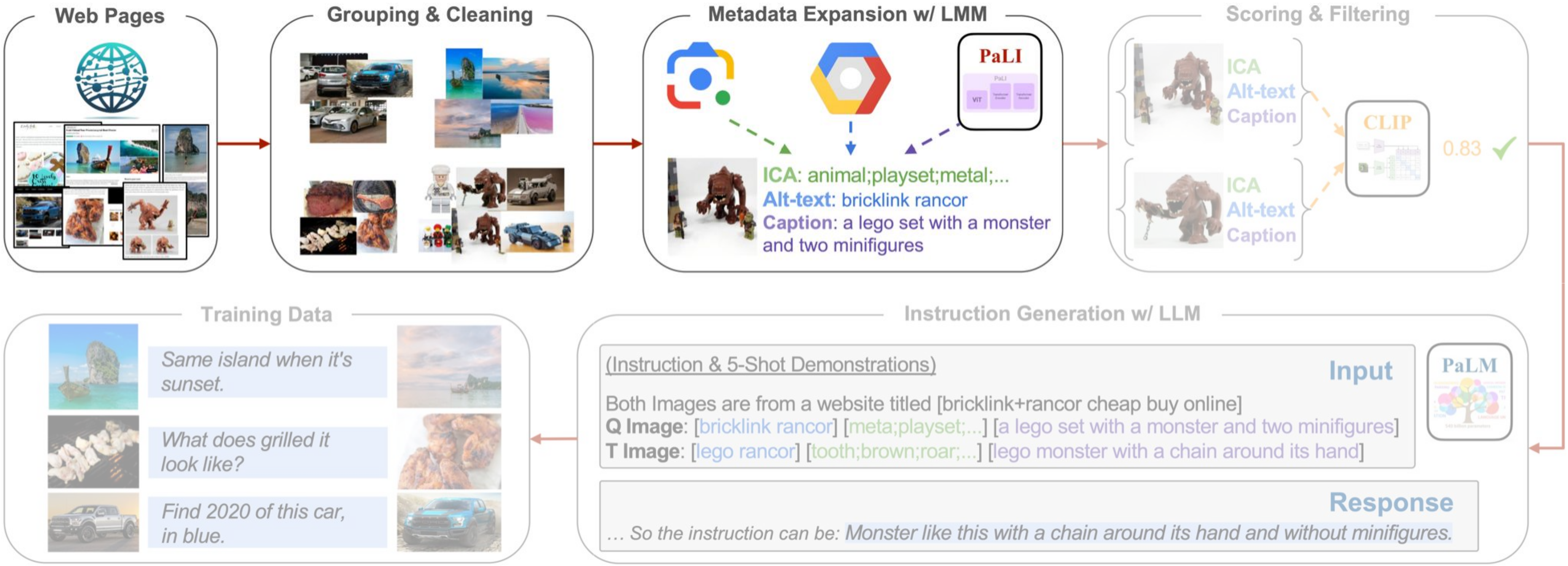
Training Data

Same island when it's sunset.

What does grilled it look like?

Find 2020 of this car, in blue.

Data Construction Pipeline



Web Pages

Grouping & Cleaning

Metadata Expansion w/ LMM

ICA: animal;playset;metal;...
Alt-text: bricklink rancor
Caption: a lego set with a monster and two minifigures

Scoring & Filtering

Instruction Generation w/ LLM

Input

(Instruction & 5-Shot Demonstrations)

Both Images are from a website titled [bricklink+rancor cheap buy online]
Q Image: [bricklink rancor] [meta;playset;...] [a lego set with a monster and two minifigures]
T Image: [lego rancor] [tooth;brown;roar;...] [lego monster with a chain around its hand]

Response

... So the instruction can be: Monster like this with a chain around its hand and without minifigures.

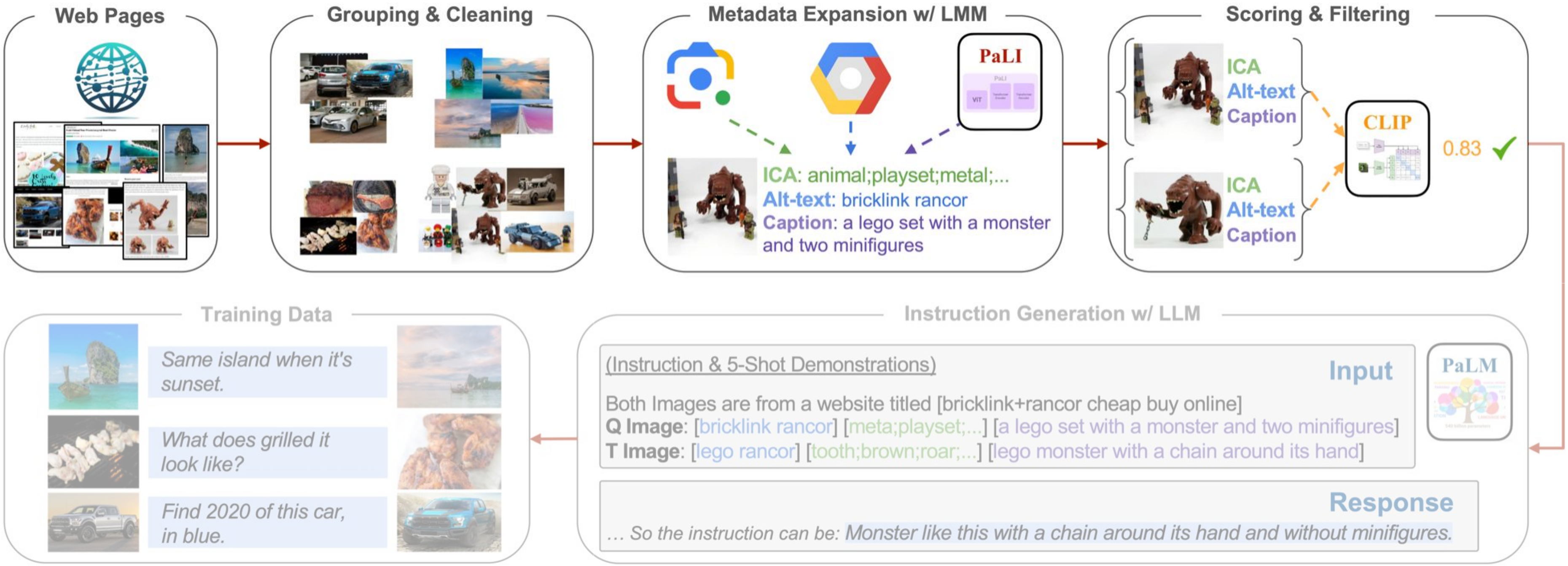
Training Data

Same island when it's sunset.

What does grilled it look like?

Find 2020 of this car, in blue.

Data Construction Pipeline



Web Pages

Grouping & Cleaning

Metadata Expansion w/ LMM

PaLI

ICA: animal;playset;metal;...
Alt-text: bricklink rancor
Caption: a lego set with a monster and two minifigures

Scoring & Filtering

ICA
Alt-text
Caption

CLIP

0.83 ✓

ICA
Alt-text
Caption

Instruction Generation w/ LLM

(Instruction & 5-Shot Demonstrations)

Input

PaLM

Both Images are from a website titled [bricklink+rancor cheap buy online]
Q Image: [bricklink rancor] [meta;playset;...] [a lego set with a monster and two minifigures]
T Image: [lego rancor] [tooth;brown;roar;...] [lego monster with a chain around its hand]

Response

... So the instruction can be: *Monster like this with a chain around its hand and without minifigures.*

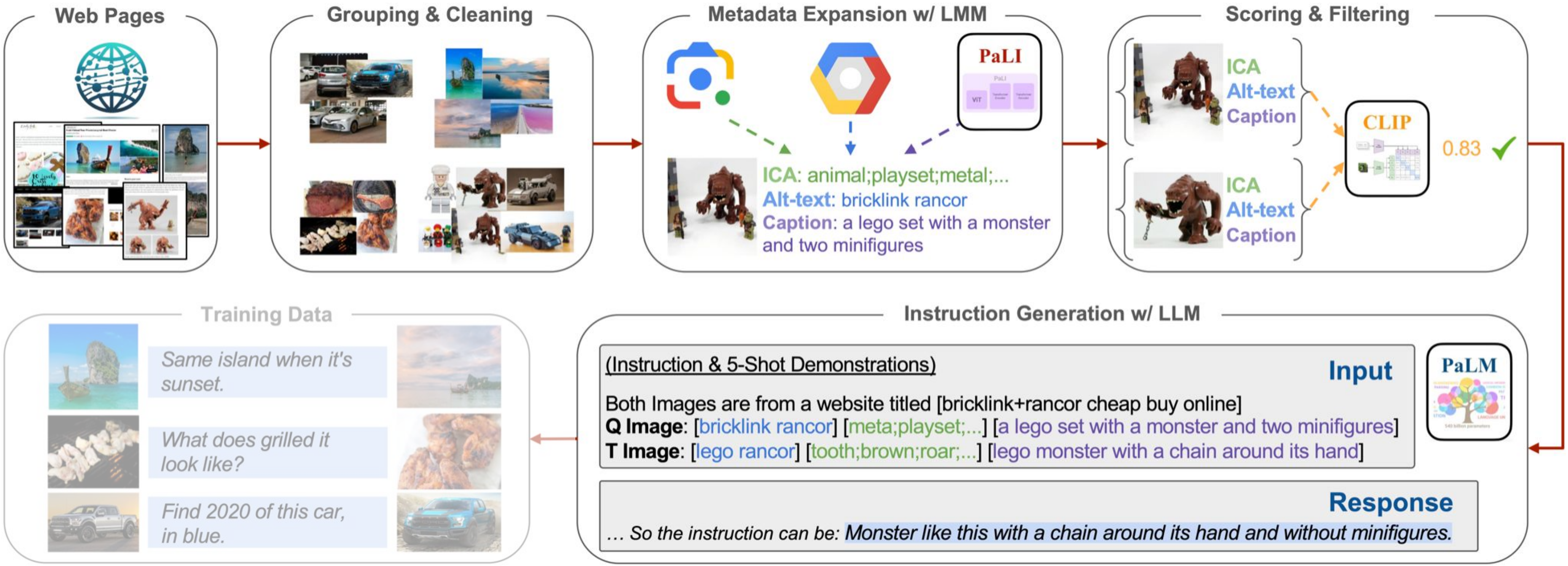
Training Data

Same island when it's sunset.

What does grilled it look like?

Find 2020 of this car, in blue.

Data Construction Pipeline



Web Pages

Grouping & Cleaning

Metadata Expansion w/ LMM

PaLI

ICA: animal;playset;metal;...
Alt-text: bricklink rancor
Caption: a lego set with a monster and two minifigures

Scoring & Filtering

ICA
Alt-text
Caption

ICA
Alt-text
Caption

CLIP

0.83 ✓

Instruction Generation w/ LLM

(Instruction & 5-Shot Demonstrations)

Input

Both Images are from a website titled [bricklink+rancor cheap buy online]
Q Image: [bricklink rancor] [meta;playset;...] [a lego set with a monster and two minifigures]
T Image: [lego rancor] [tooth;brown;roar;...] [lego monster with a chain around its hand]

PaLM

Response

… So the instruction can be: *Monster like this with a chain around its hand and without minifigures.*

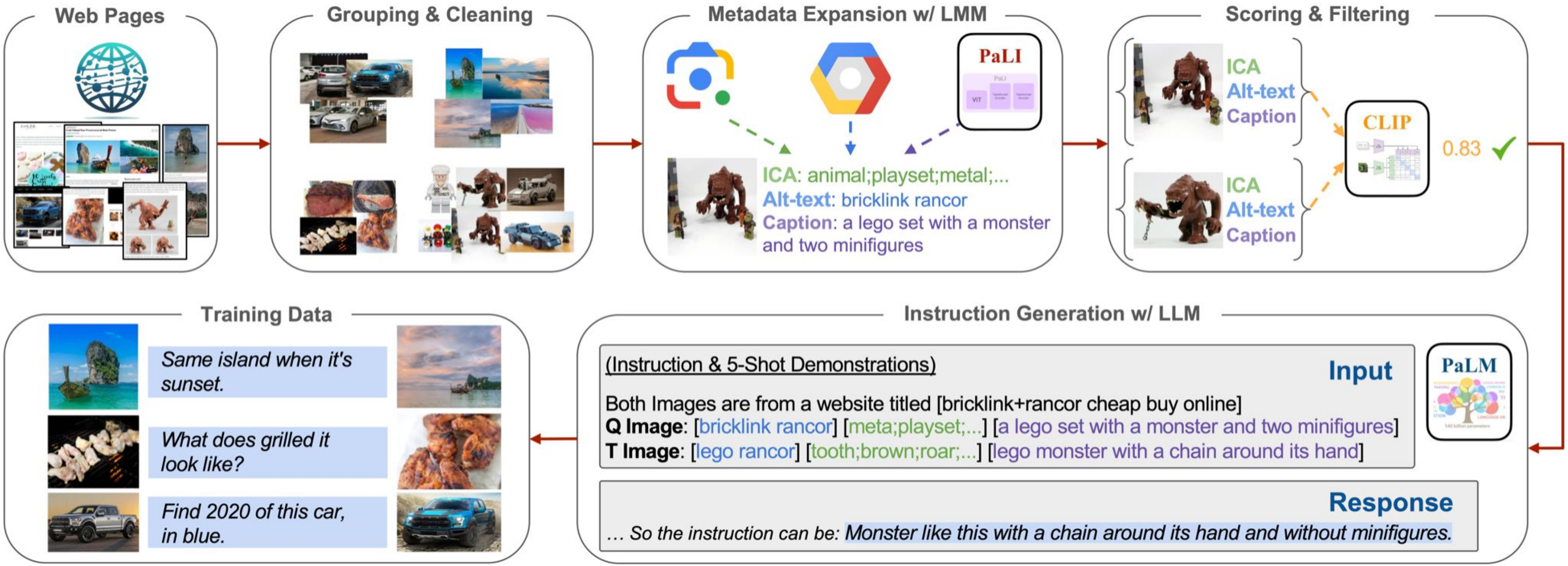
Training Data

Same island when it's sunset.

What does grilled it look like?

Find 2020 of this car, in blue.

Data Construction Pipeline



Web Pages

Grouping & Cleaning

Metadata Expansion w/ LMM

ICA: animal;playset;metal;...
Alt-text: bricklink rancor
Caption: a lego set with a monster and two minifigures

Scoring & Filtering

Instruction Generation w/ LLM

(Instruction & 5-Shot Demonstrations) **Input**

Both Images are from a website titled [bricklink+rancor cheap buy online]
Q Image: [bricklink rancor] [meta;playset;...] [a lego set with a monster and two minifigures]
T Image: [lego rancor] [tooth;brown;roar;...] [lego monster with a chain around its hand]

Response

… So the instruction can be: Monster like this with a chain around its hand and without minifigures.

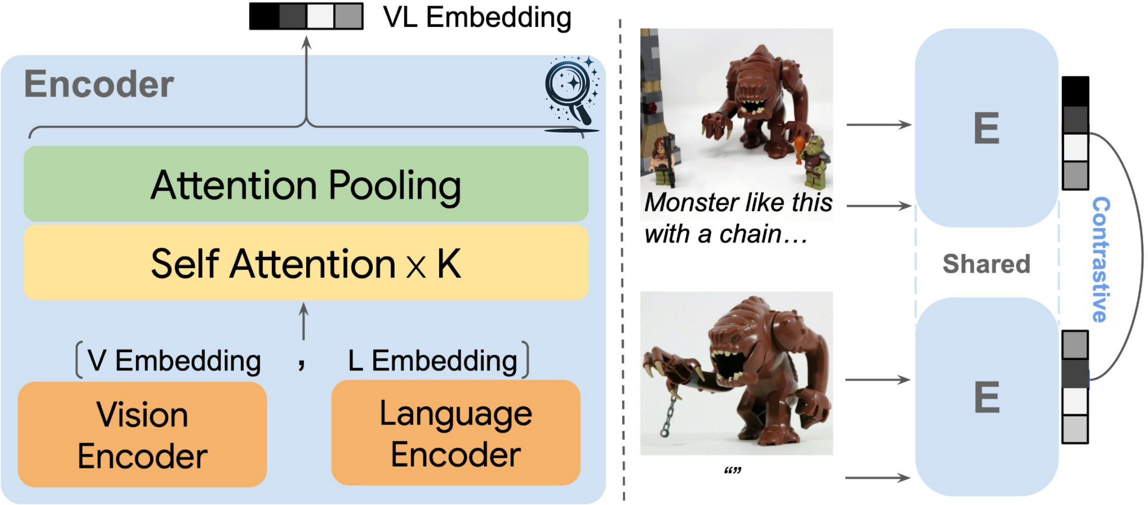
Training Data

Same island when it's sunset.

What does grilled it look like?

Find 2020 of this car, in blue.

Training Model with Simple Contrastive Loss

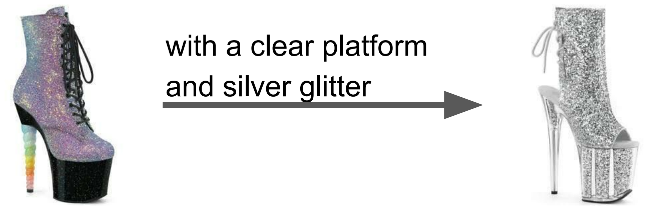


$$-\log \frac{e^{\mathrm{sim}(\boldsymbol{r}_q^i,\boldsymbol{r}_t^i)/\tau}}{\sum_{j=1}^N \left(e^{\mathrm{sim}(\boldsymbol{r}_q^i,\boldsymbol{r}_t^j)/\tau} + \boxed{e^{\mathrm{sim}(\boldsymbol{r}_q^i,\boldsymbol{r}_t^{j'})/\tau}}\right)},$$

Query Negative

Evaluation - Multimodality to Image Retrieval

Composed Image Retrieval



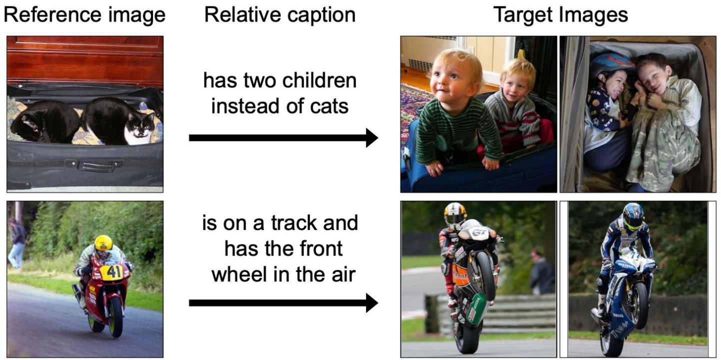
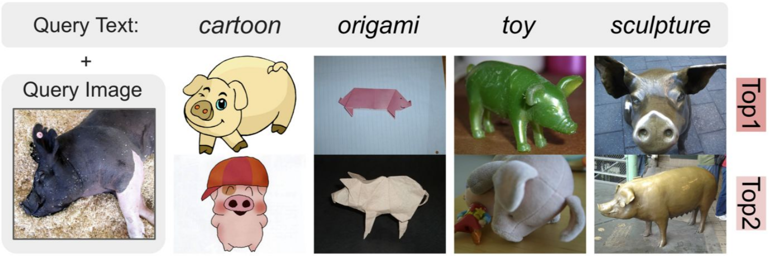
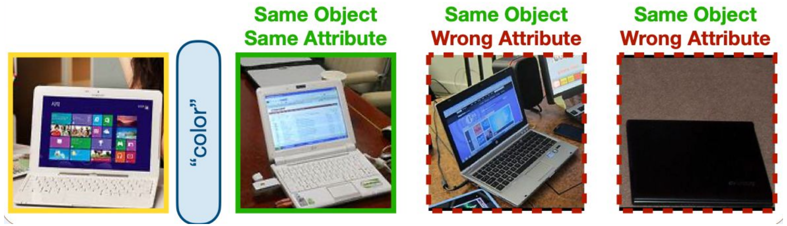


Figure 3. Examples of CIR queries and ground truths in CIRCO.

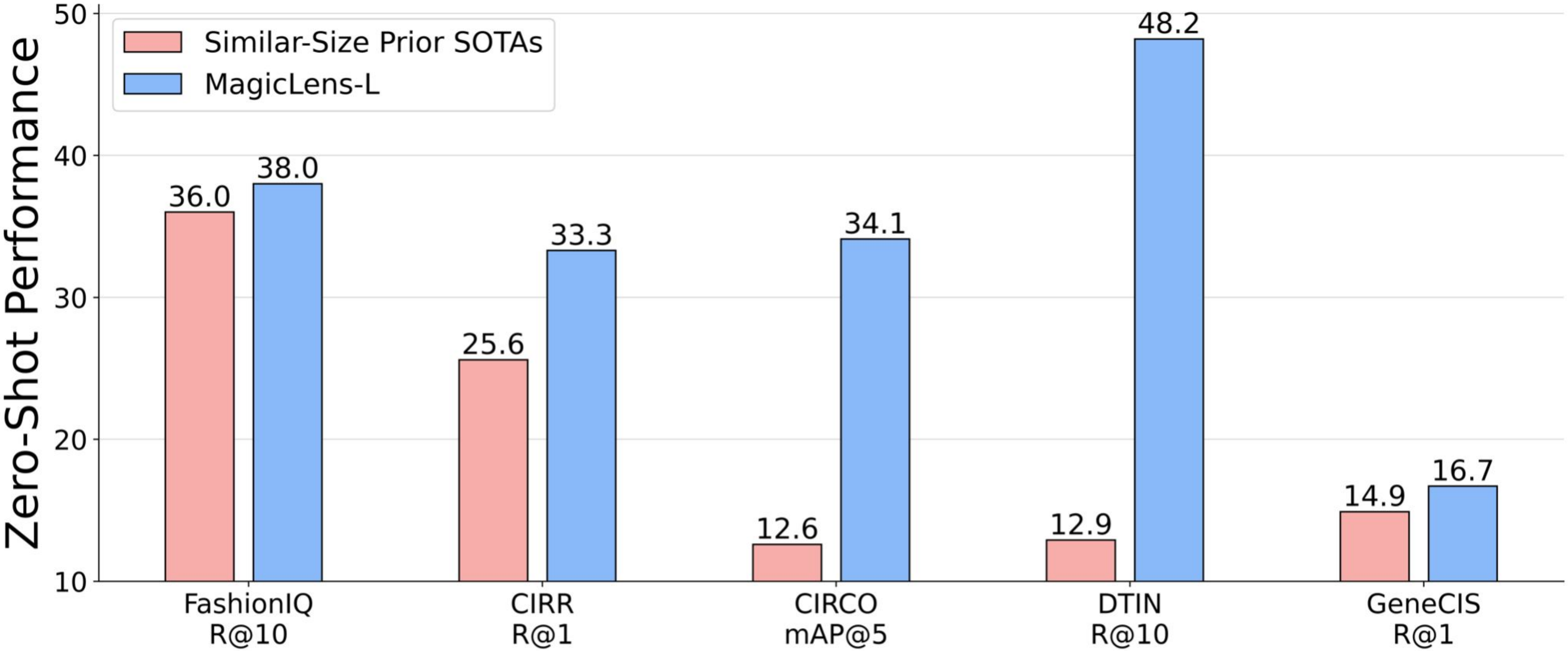
Domain Transfer Retrieval



Conditional Similarity

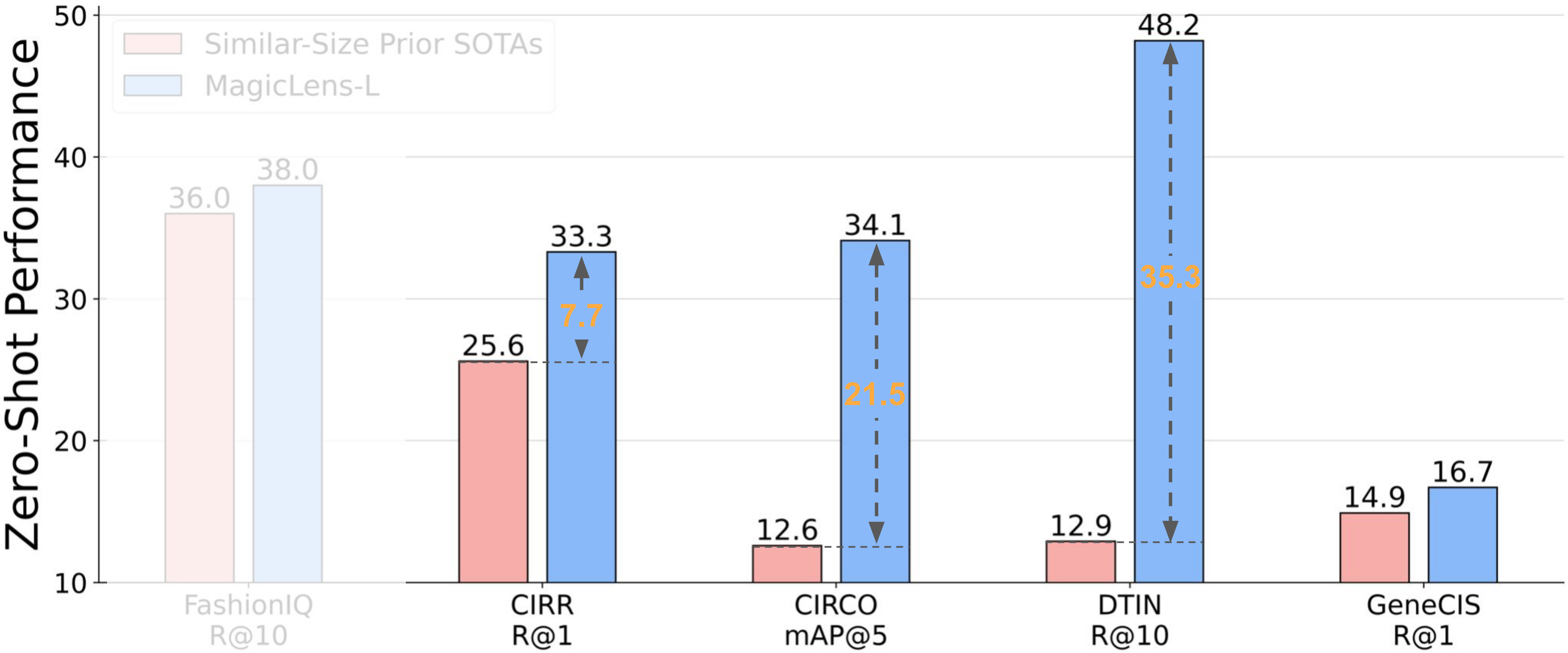


Main Results (Multimodality-to-Image Retrieval)



SOTA on **5** benchmarks of **3** multimodality-to-image retrieval tasks.

Main Results (Multimodality-to-Image Retrieval)



Significant improvements on open-domain images.

Image-to-Image Retrieval

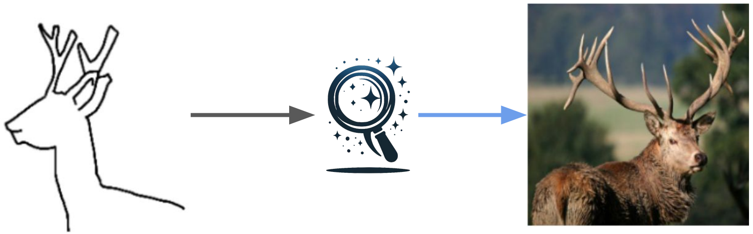
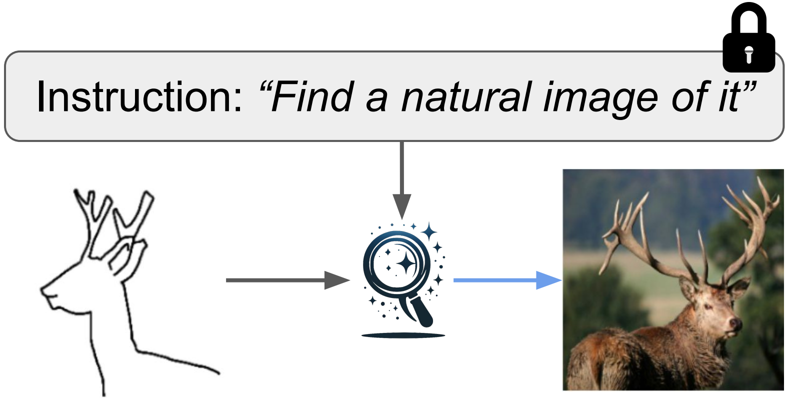
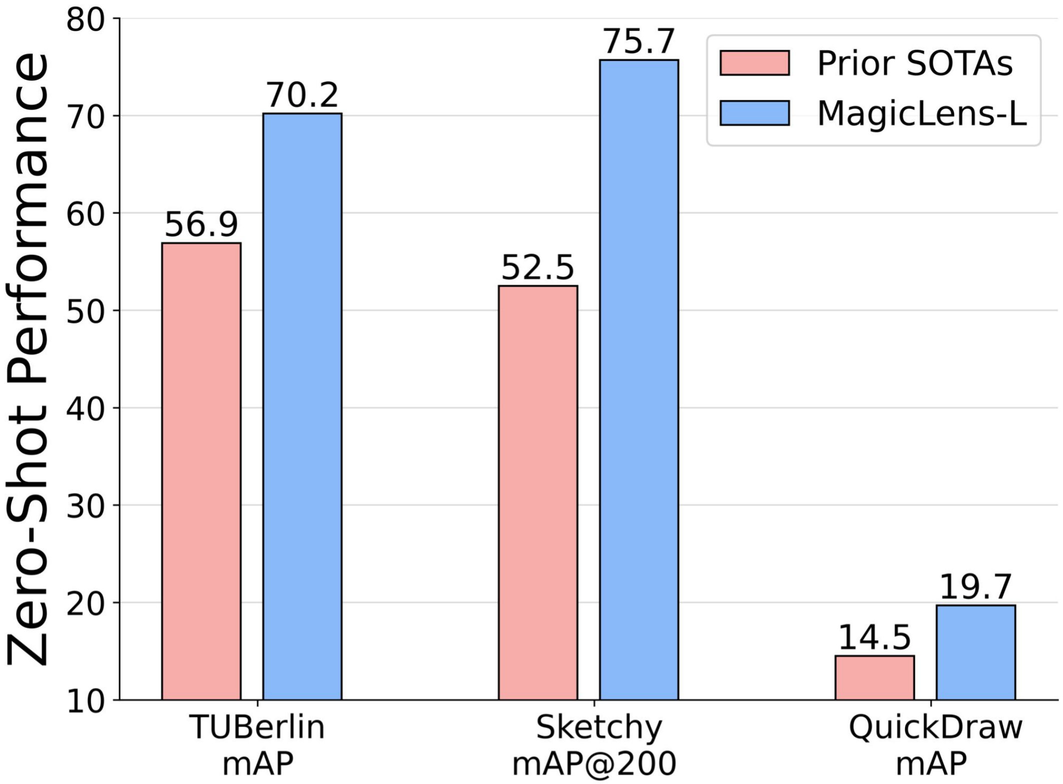


Image-to-Image Retrieval

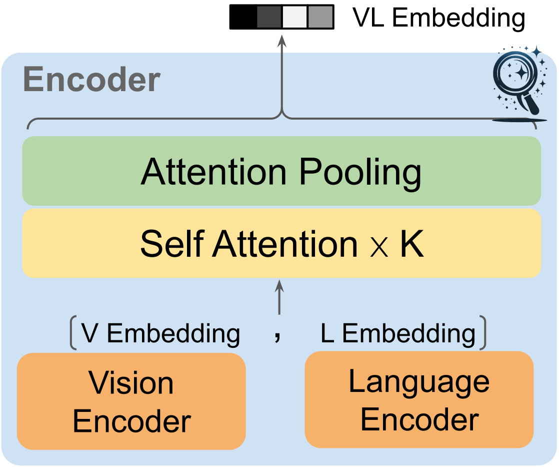


Instruction: *“Find a natural image of it”*

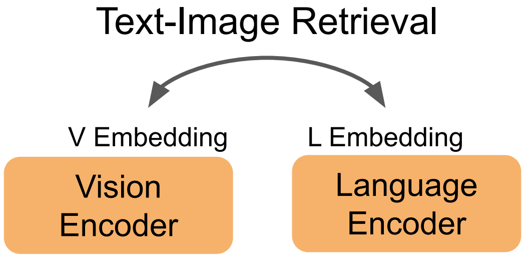


Without fine-tuning, a **single** MagicLens model outperforms **fine-tuned** SOTAs.

Text-to-Image Retrieval



Text-to-Image Retrieval over Flickr and COCO



Text-Image Retrieval

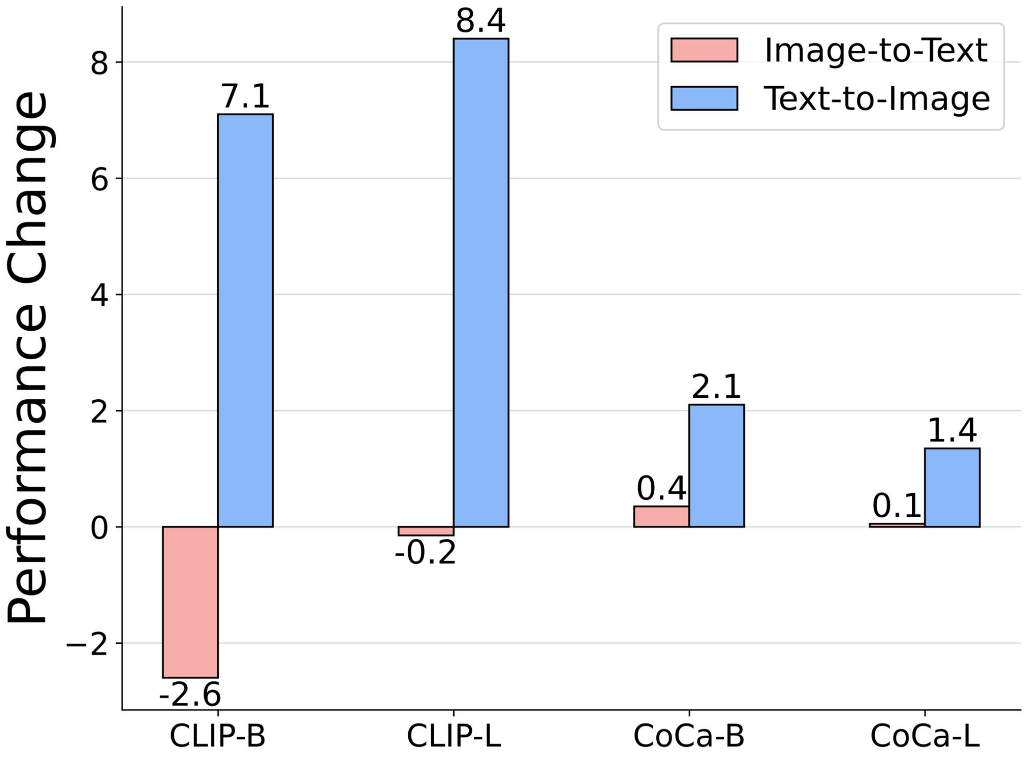
V Embedding

L Embedding

Vision Encoder

Language Encoder

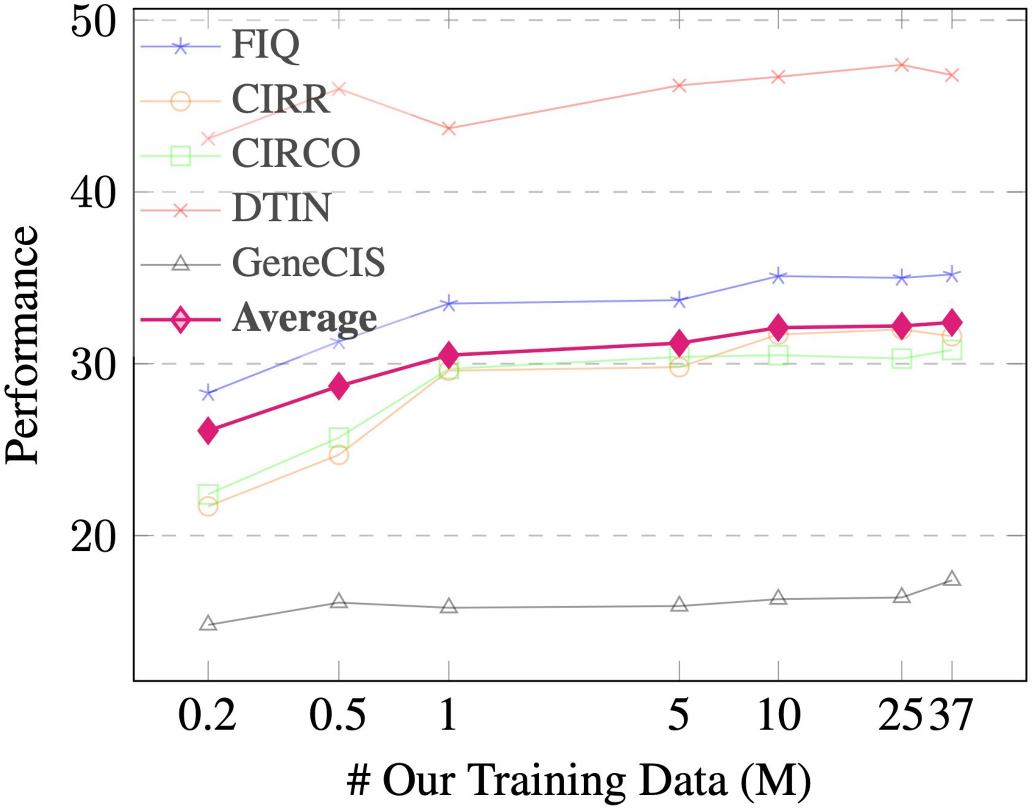
1. No or marginal drop on I2T
2. Moderate to **significant** Improvements on T2I
3. Non-trivial overall Improvements



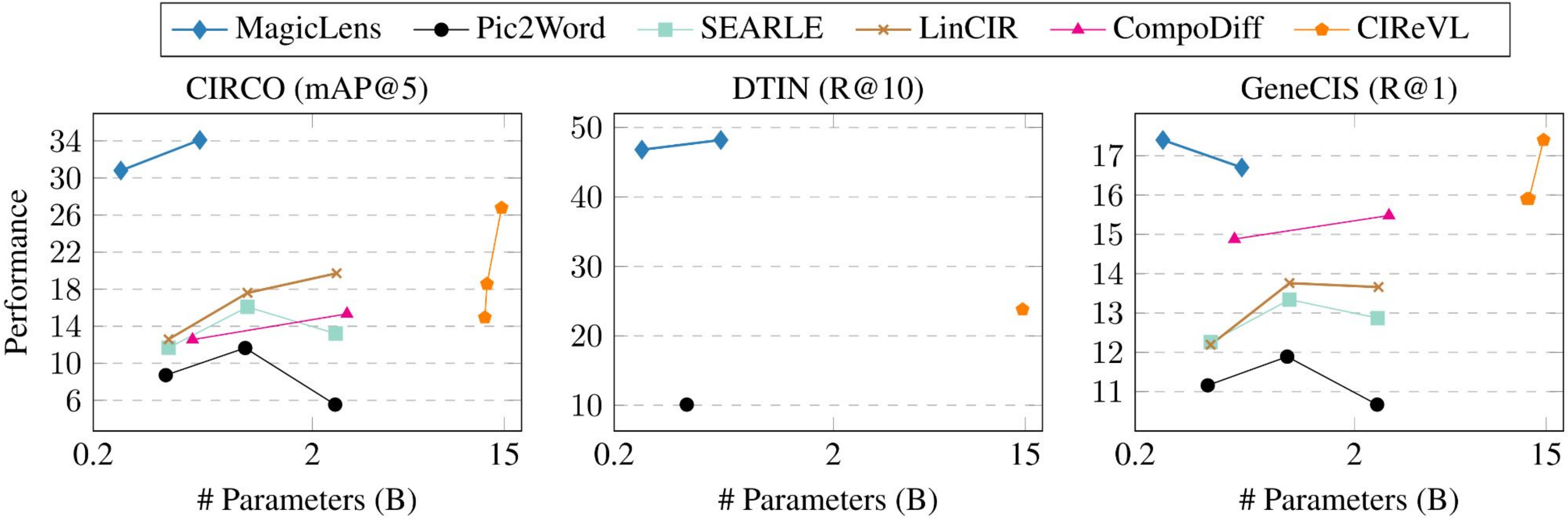
Data Analysis - Scaling

1. More data, stronger model

2. SOTA with 1M data



Model Analysis - Parameter Efficiency

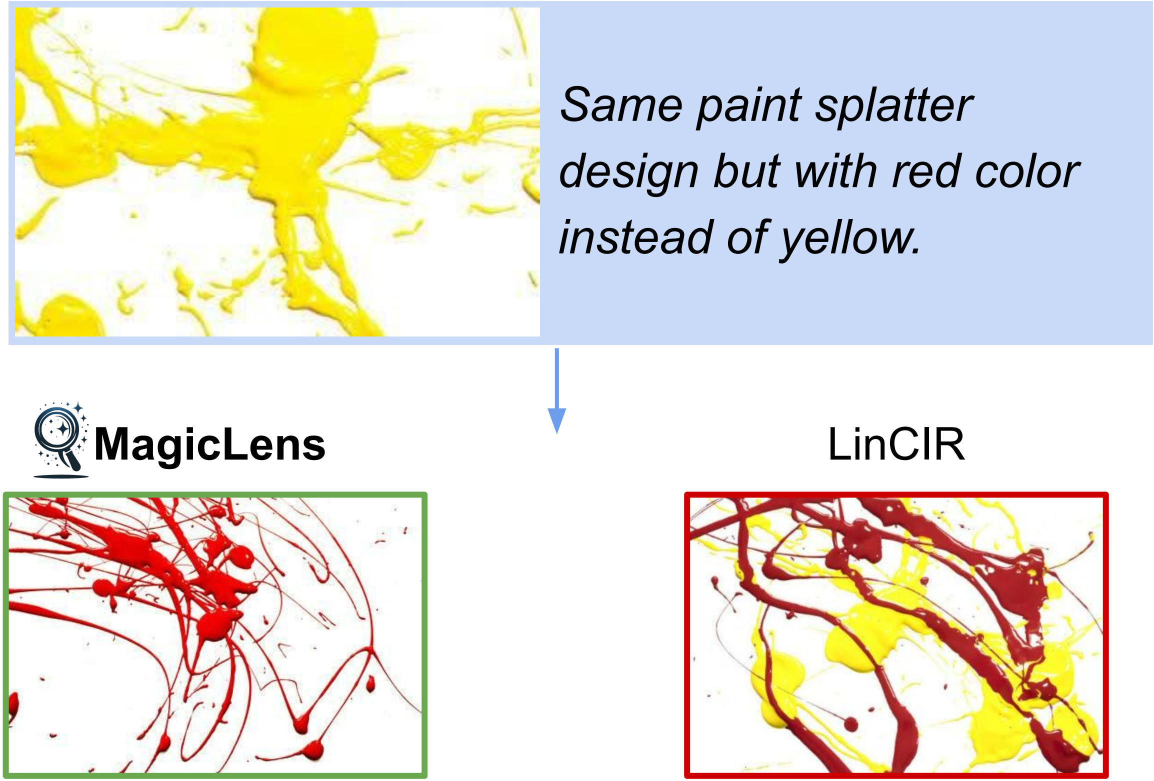


MagicLens outperforms SOTA on three tasks with **50X smaller** #Params.

Analysis on 1.4M Image Pool

Simple Visual

Instruction	MagicLens-L	LinCIR	Tie
Simple Visual	**50.7**	41.3	8.0



Same paint splatter design but with red color instead of yellow.

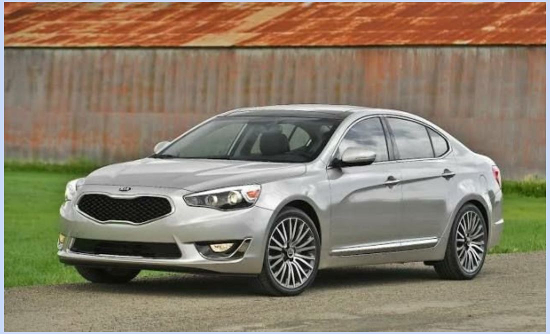
MagicLens

LinCIR

Analysis on 1.4M Image Pool

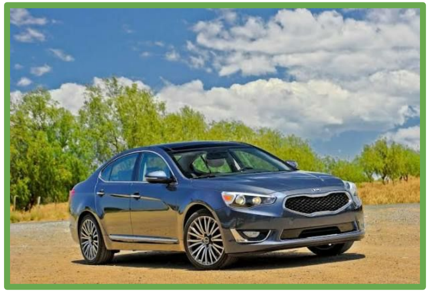
Complex Visual

Instruction	MagicLens-L	LinCIR	Tie
Simple Visual	**50.7**	41.3	8.0
Complex Visual	**61.3**	24.0	14.7

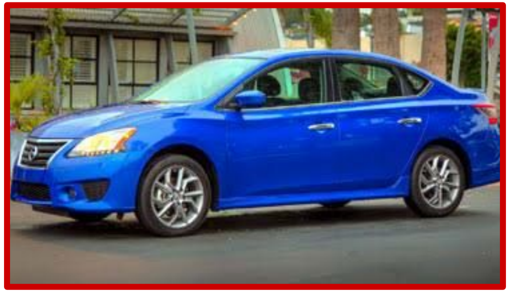


Same car model as the given image, but 1) a 2013 model, 2) blue in color, and 3) parked in front of trees.

MagicLens



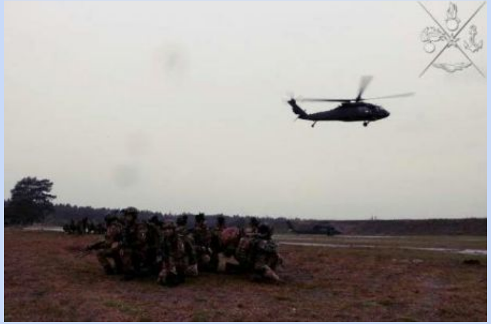
LinCIR



Analysis on 1.4M Image Pool

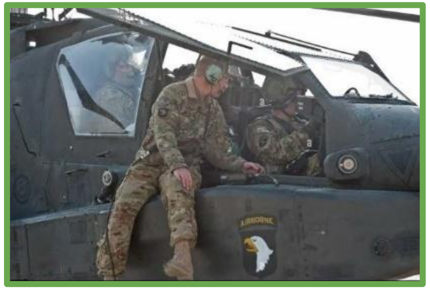
Beyond Visual

Instruction	MagicLens-L	LinCIR	Tie
Simple Visual	**50.7**	41.3	8.0
Complex Visual	**61.3**	24.0	14.7
Beyond Visual	**80.0**	4.7	15.3

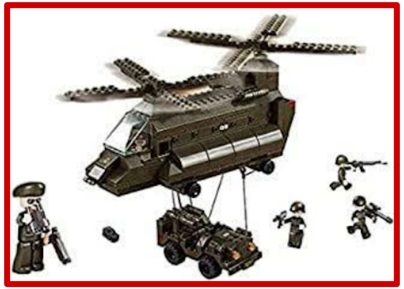


landed and the soldiers sitting in front of it.

MagicLens



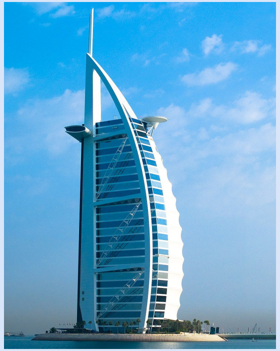
LinCIR



MagicLens handles all three search intents, especially **complex** and **beyond visual** ones.

Google DeepMind

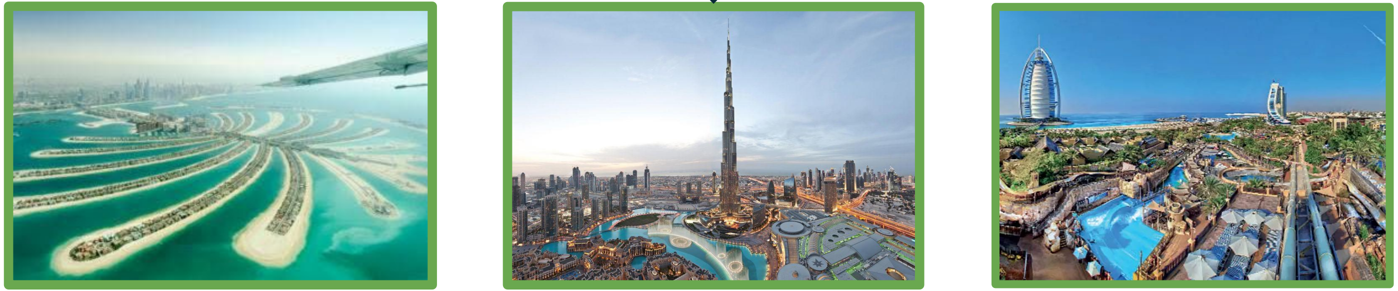
NLP



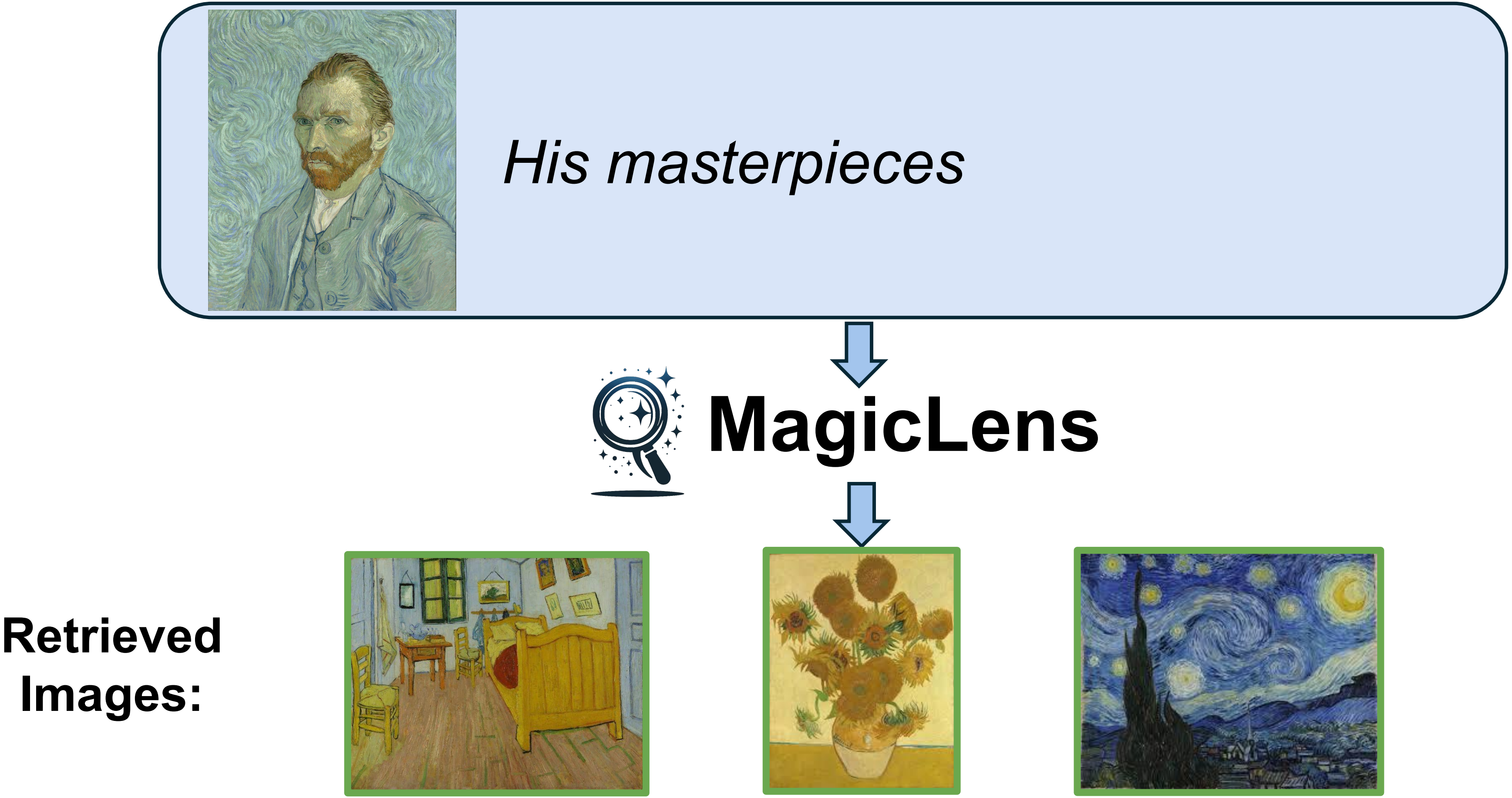
Find other attractions in this country

MagicLens

Retrieved Images:



Beyond Visual Similarity

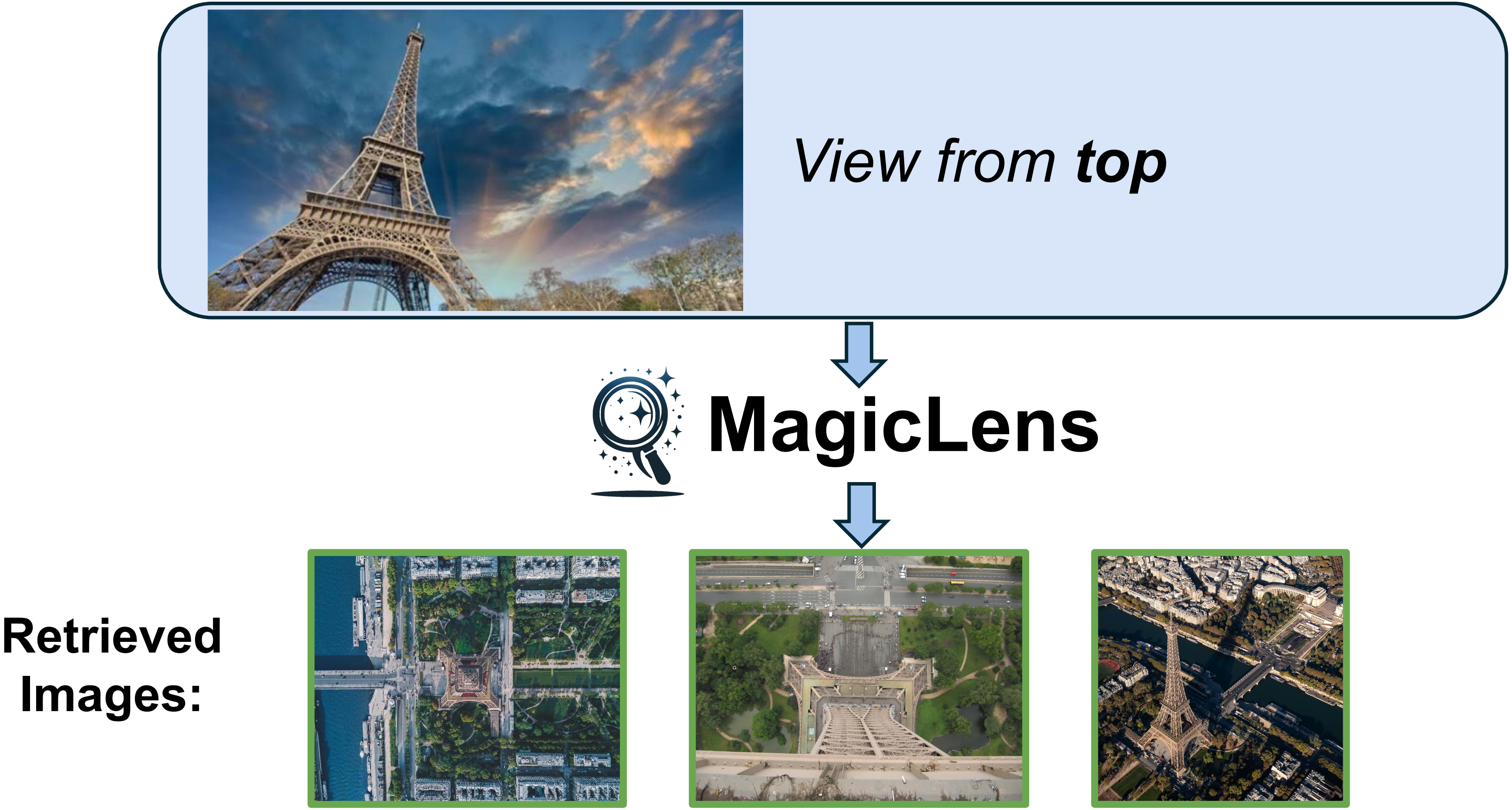


His masterpieces

MagicLens

Retrieved Images:

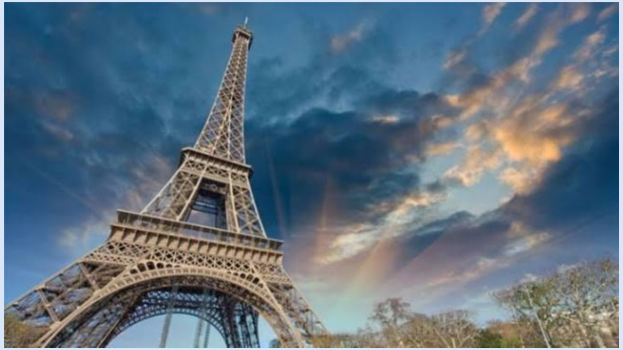
Beyond Visual Similarity



*View from **top***

MagicLens

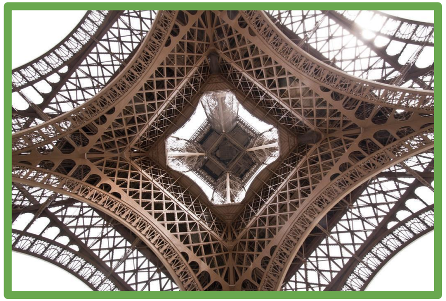
Retrieved Images:

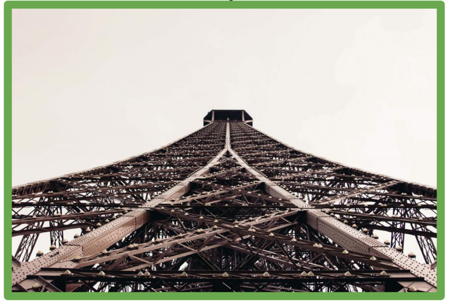


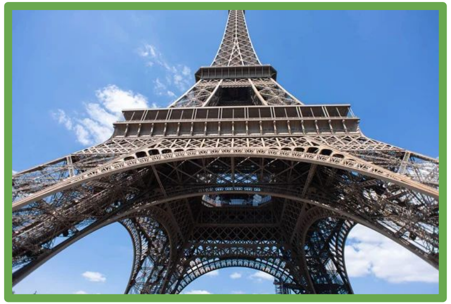
*View from **bottom***

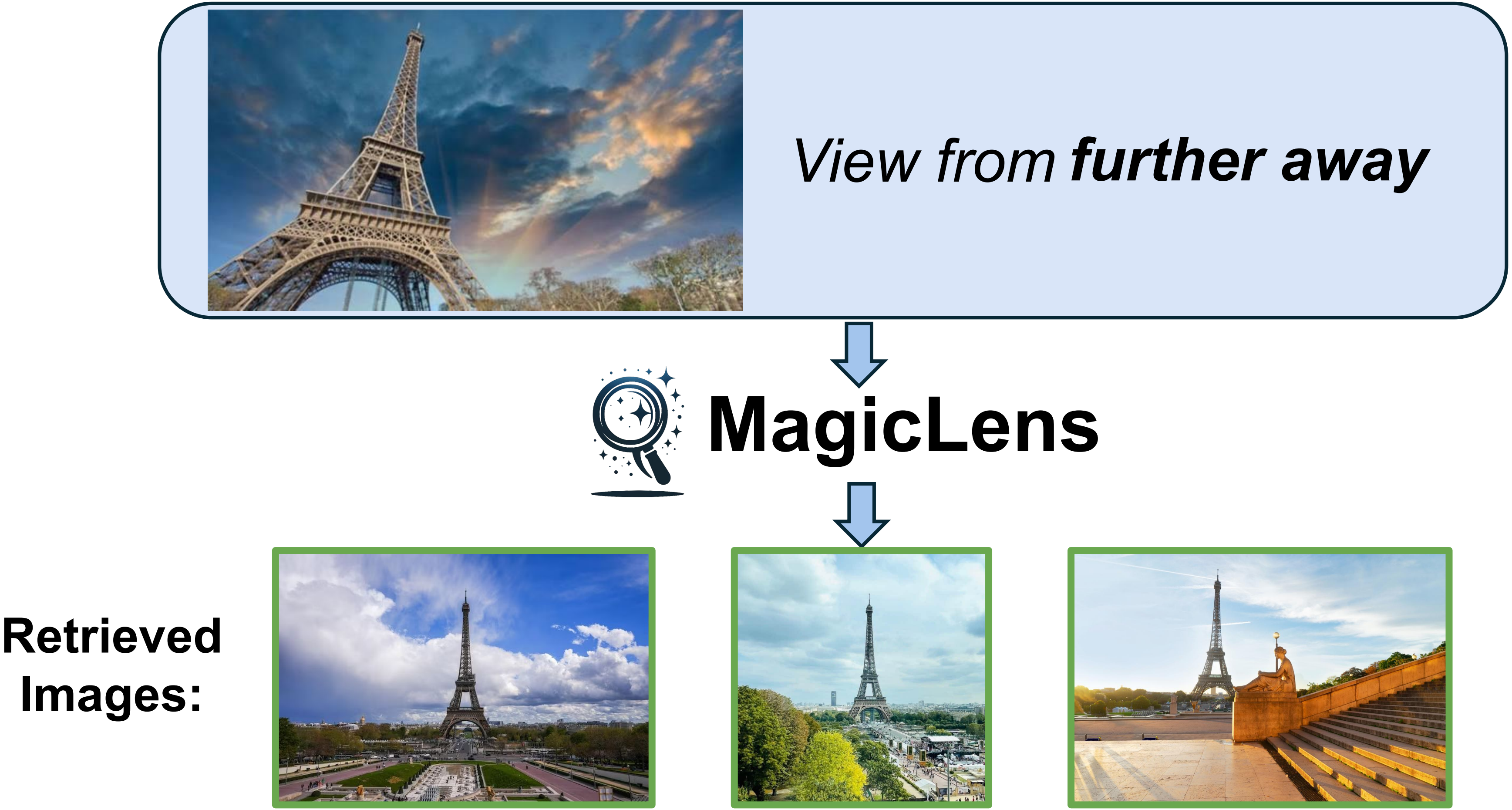
MagicLens

Retrieved Images:









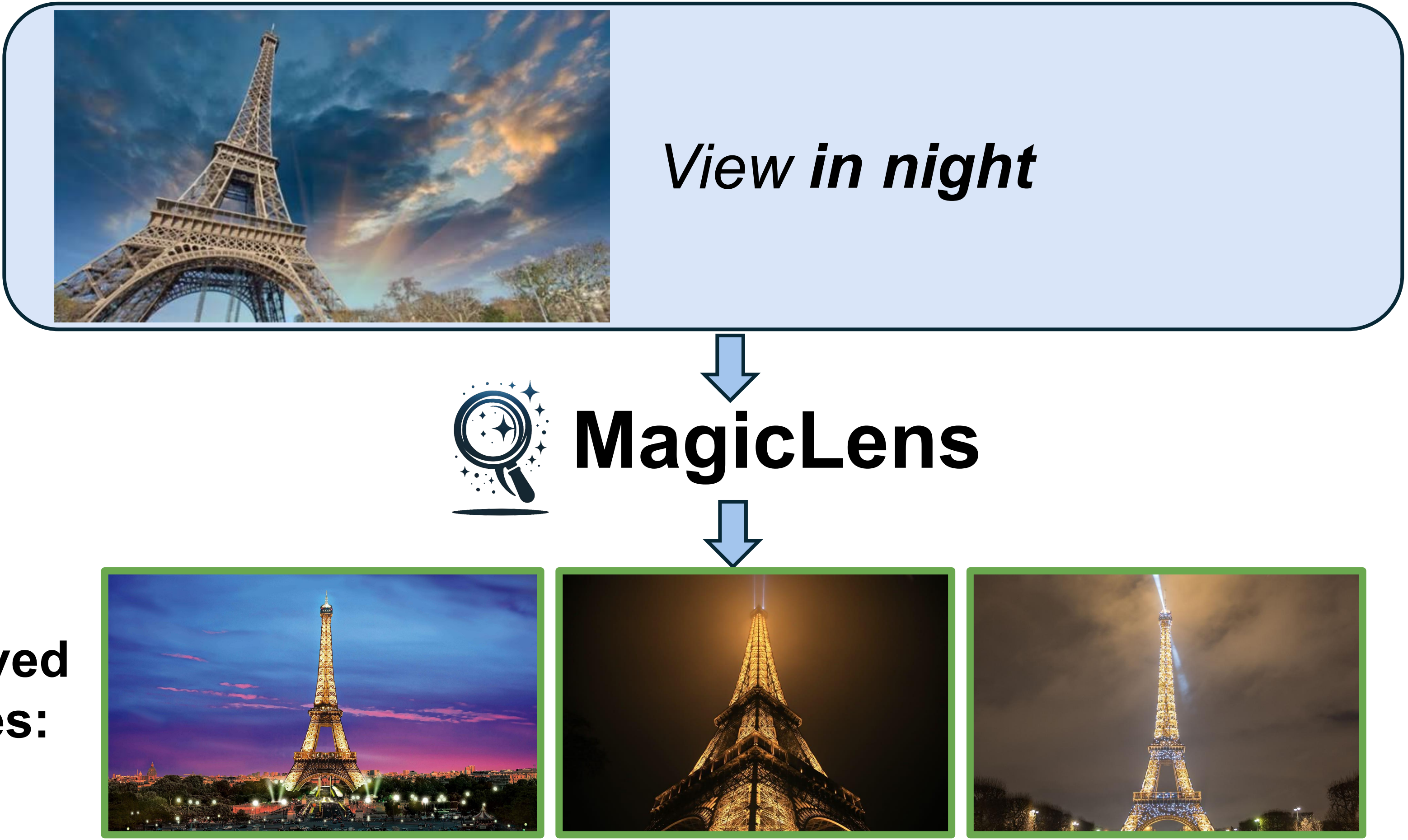
*View from **further away***

MagicLens

Retrieved Images:

Google DeepMind

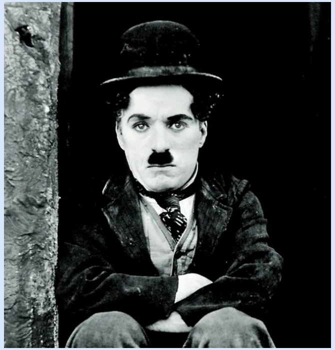
NLP



View *in night*

MagicLens

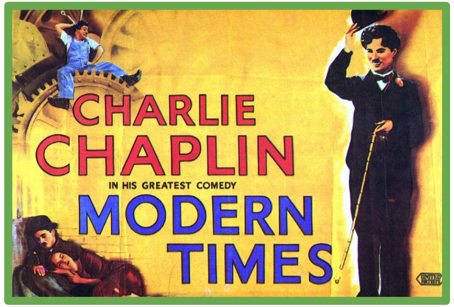
Retrieved Images:

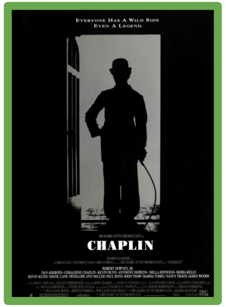


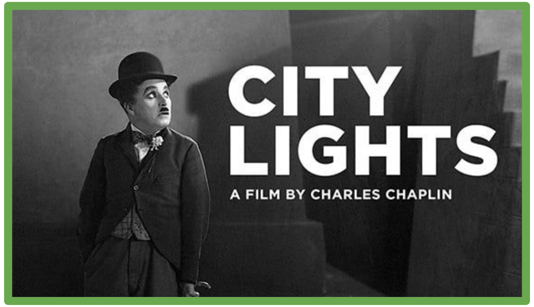
Recommend some movies he is in

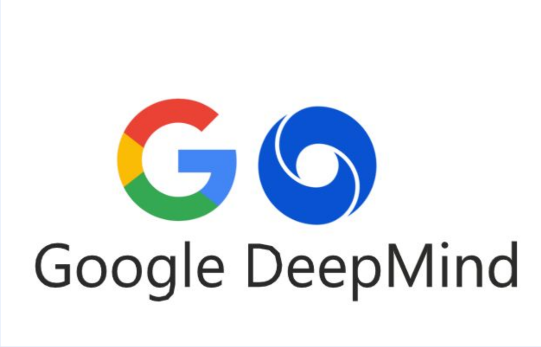
MagicLens

Retrieved Images:





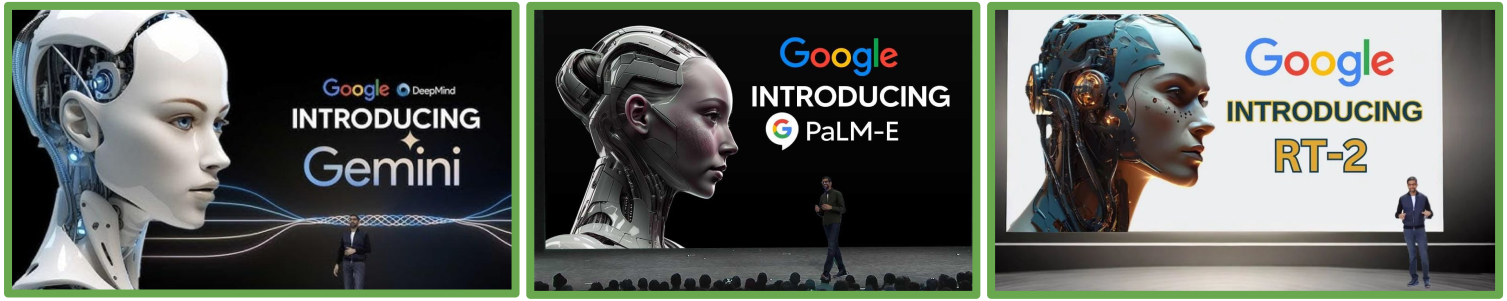




What's the new product?

MagicLens

Retrieved Images:



Conclusions

- MagicLens is a series retrieval models that are:
  - **lightweight** (50X smaller than prior SOTAs)
  - **powerful** (strong results across 10 benchmarks)
  - **unified** (multimodal-, image-, and text-to-image retrieval)
  - **open-ended** (satisfying open-ended search intents)

Code&Model

Paper

Project Page