

# A Comprehensive Study of Multimodal Large Language Models for Image Quality Assessment

Tianhe Wu<sup>1,2</sup>, Kede Ma<sup>2\*</sup>, Jie Liang<sup>3</sup>, Yujiu Yang<sup>1\*</sup>, Lei Zhang<sup>3,4</sup>

<sup>1</sup>Tsinghua University

<sup>2</sup>Department of Computer Science, City University of Hong Kong

<sup>3</sup>OPPO Research Institute

<sup>4</sup>Department of Computing, The Hong Kong Polytechnic University

<https://github.com/TianheWu/MLLMs-for-IQA>



清華大學

Tsinghua University



香港城市大學  
City University of Hong Kong

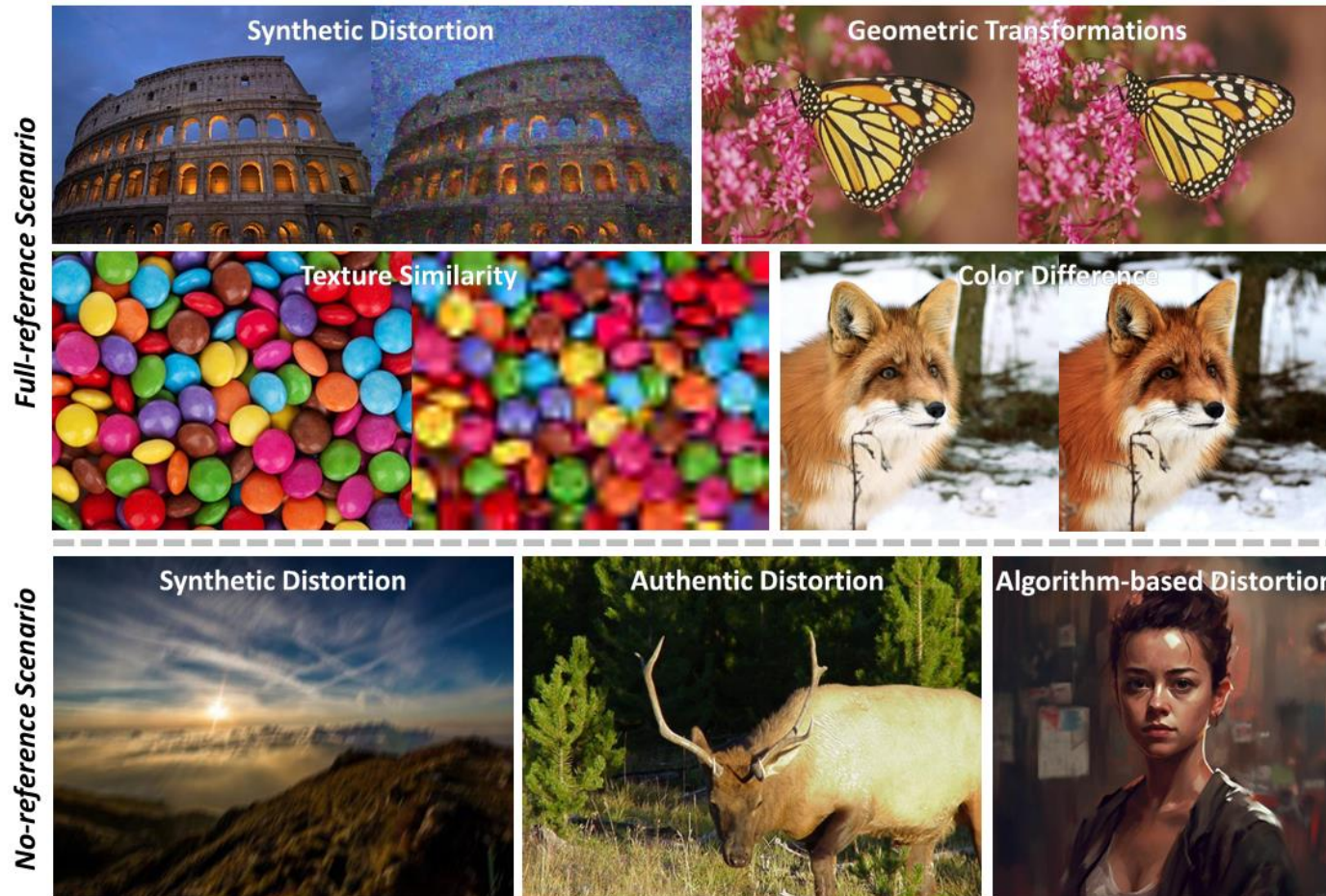
oppo



THE HONG KONG  
POLYTECHNIC UNIVERSITY  
香港理工大學

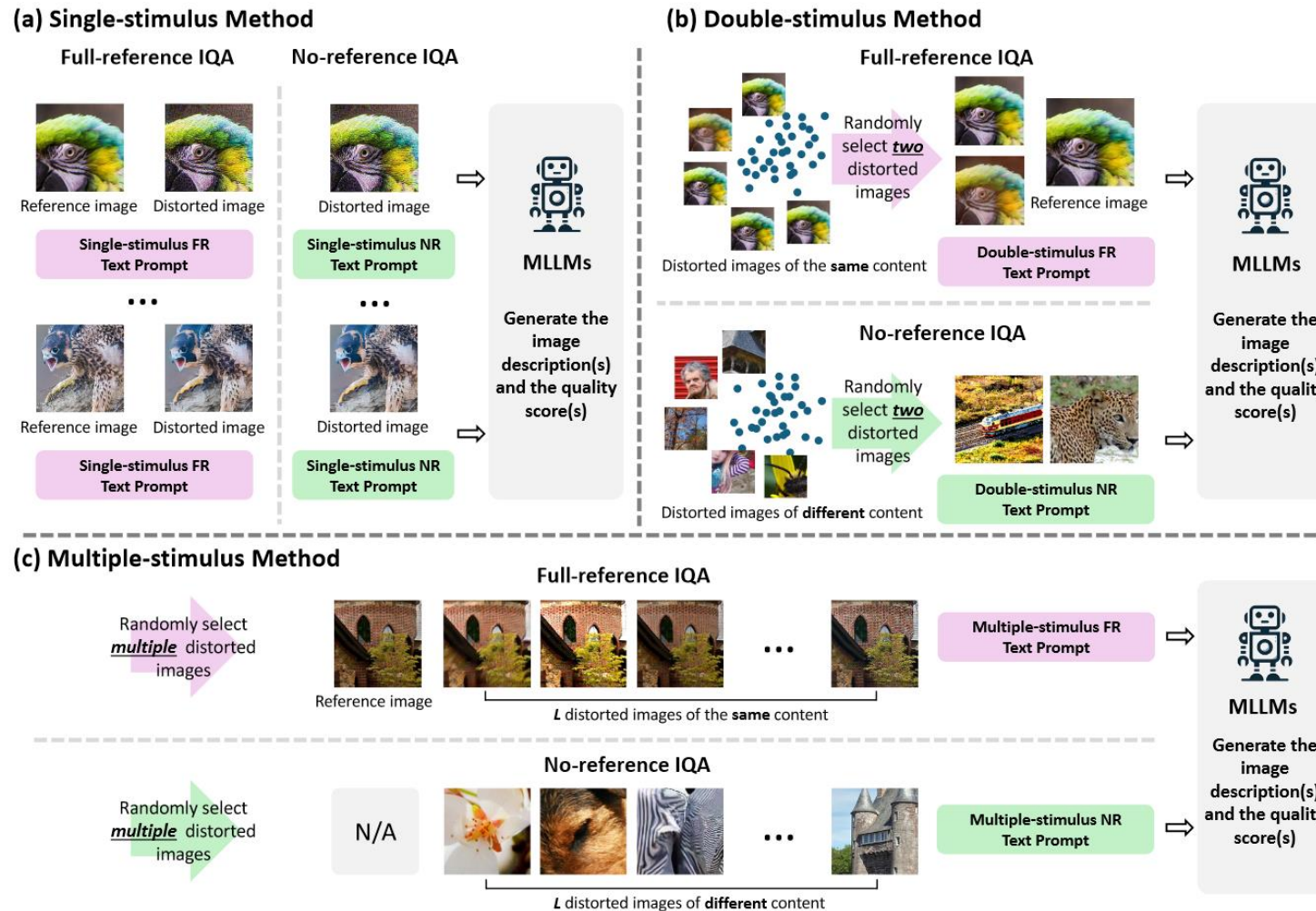
# Motivation

- We aim to evaluate the capability of MLLMs on Image Quality Assessment (IQA) task
- We conduct a comprehensive and systematic study of **prompting** MLLMs for IQA



# Prompting MLLMs for IQA

- Prompting Strategies from Psychophysics



- Single-stimulus method
- Double-stimulus method
- Multiple-stimulus method



# Prompting MLLMs for IQA

- Prompting Strategies from NLP

## (a) Standard Prompting

### Single-stimulus Method



Please assign a perceptual quality score in terms of [...]. The score must range from 0 to 100, with a higher score denoting better image quality. [...]

### Double-stimulus Method



Please assign a perceptual quality comparison result between the two images in terms of [...]. If you judge that the first image has better quality than the second image, output 1, if you judge that the second image has better quality than the first image, output 0, if you judge that two images have the same quality, output 2. [...]

### Multiple-stimulus Method



Please assign a perceptual quality ranking result among four images in terms of [...]. The image with the lowest perceptual quality is ranked 0, and the image with the highest perceptual quality is ranked 3. If you judge that some distorted images have the same perceptual quality, their ranking can be the same. [...]

## (b) Chain-of-thought Prompting

### Single-stimulus Method



Please first detail its perceptual quality in terms of [...]. Then, based on the perceptual analysis of the given image, assign a quality score to the given image. The score must range from 0 to 100, with a higher score denoting better image quality. [...]

### Double-stimulus Method



Please first detail their perceptual quality comparison in terms of [...]. Then, based on the quality comparison analysis between them, assign a perceptual quality comparison result between the two images. If you judge that the first image has better quality than the second image, output 1, if you judge that the second image has better quality than the first image, output 0, if you judge that two images have the same quality, output 2. [...]

### Multiple-stimulus Method



Please first detail their perceptual quality comparison in terms of [...]. Then, based on the quality comparison analysis among them, please assign a perceptual quality ranking result among four images. The image with the lowest perceptual quality is ranked 0, and the image with the highest perceptual quality is ranked 3. If you judge that some distorted images have the same perceptual quality, their ranking can be the same. [...]

## (c) In-context Prompting

### Single-stimulus Method



For the shown two images, the human perceptual quality score of the first image is 50. Now, based on the above example, please assign a perceptual quality score to the second image in terms of [...]. The score must range from 0 to 100, with a higher score denoting better image quality. [...]

### Double-stimulus Method



For the first two images (the first and the second images), the human perceptual quality comparison result is that the first image is of better quality than the second image. Now, based on the above example, please assign a perceptual quality comparison result between the second two images (the third and the fourth images) in terms of [...]. If you judge that the third image has better quality than the fourth image, output 1, if you judge that the fourth image has better quality than the third image, output 0, if you judge that two images have the same quality, output 2. [...]

### Multiple-stimulus Method



For the shown eight images, for the first four images (from the first to the fourth images), the human perceptual quality ranking result is [first: 0, second: 1, third: 2, fourth: 3]. Now, based on the above example, please assign a perceptual quality ranking result among the second four images (from the fifth to the eighth images) in terms of [...]. The image with the lowest perceptual quality is ranked 0, and the image with the highest perceptual quality is ranked 3. If you judge that some distorted images have the same perceptual quality, their ranking can be the same. [...]

- Standard prompting
- Chain-of-thought prompting
- In-context prompting

# Computational Procedure for Difficult Sample Selection



EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO  
2024

- Computational Procedure for Difficult Sample Selection
  - **Inference with MLLMs tends to be slow and costly**
  - First, they should be **difficult**, with a high likelihood of causing MLLMs to err
  - Second, the selected samples should be **diverse** to highlight various aspects of MLLMs' potential weaknesses in IQA
  - Third, they should be **consistent** with small variations in individual quality scores

- FR Scenario:

$$y_n^* = \operatorname{argmax}_{y \in \mathcal{Y} \setminus \mathcal{Y}_{n-1}^*} \frac{1}{|\mathcal{X}_y|} \sum_{x \in \mathcal{X}_y} \frac{(d_w(x, y) - q(x))^2}{(\sigma(x))^2 + \epsilon} + \lambda \operatorname{Div}(y, \mathcal{Y}_{n-1}^*), \quad x_k^* = \operatorname{argmax}_{x \in \mathcal{X}_y \setminus \mathcal{X}_{k-1}^*} \frac{(d_w(x, y) - q(x))^2}{(\sigma(x))^2 + \epsilon}, \quad y \in \mathcal{Y}_N^*,$$

- NR scenario:

$$x_n^* = \operatorname{argmax}_{x \in \mathcal{X} \setminus \mathcal{X}_{n-1}^*} \frac{(q_w(x) - q(x))^2}{(\sigma(x))^2 + \epsilon} + \lambda \operatorname{Div}(x, \mathcal{X}_{n-1}^*),$$

# Comparison of Nine Prompting Systems



Method	FR IQA				NR IQA		
	FR-KADID	Aug-KADID	TQD	SPCD	NR-KADID	SPAQ	AGIQA-3K
Single-stimulus Method							
LLaVA-v1.6-S	0.227	0.013	0.180	0.001	0.262	0.544	<b>0.614</b>
mPLUG-Owl2-S	0.285	0.218	0.228	0.081	0.126	0.467	0.279
InternLM-XC2-VL-S	0.274	0.272	0.299	0.009	0.252	0.794	0.512
GPT-4V-S	<b>0.745</b>	<b>0.786</b>	<b>0.773</b>	<b>0.098</b>	<b>0.467</b>	<b>0.860</b>	0.420
LLaVA-v1.6-I	0.249	0.194	0.222	<b>0.147</b>	0.116	0.019	0.061
mPLUG-Owl2-I	0.373	0.373	0.246	0.047	0.017	0.083	0.409
InternLM-XC2-VL-I	0.380	0.241	0.204	0.087	0.188	0.342	0.461
GPT-4V-I	<b>0.771</b>	<b>0.753</b>	<b>0.738</b>	0.028	<b>0.590</b>	<b>0.845</b>	<b>0.650</b>
LLaVA-v1.6-C	0.164	0.300	0.226	<b>0.174</b>	0.151	0.550	0.580
mPLUG-Owl2-C	0.387	0.361	0.278	0.122	0.179	0.455	0.409
InternLM-XC2-VL-C	0.237	0.306	0.167	0.063	0.306	0.649	0.507
GPT-4V-C	<b>0.809</b>	<b>0.782</b>	<b>0.809</b>	0.121	<b>0.517</b>	<b>0.869</b>	<b>0.677</b>
Double-stimulus Method							
LLaVA-v1.6-S	0.387	0.396	0.390	0.113	0.270	0.430	0.234
mPLUG-Owl2-S	0.435	0.307	0.350	<b>0.117</b>	0.126	0.157	0.020
InternLM-XC2-VL-S	0.309	0.408	0.440	0.042	0.267	0.690	0.555
GPT-4V-S	<b>0.679</b>	<b>0.743</b>	<b>0.655</b>	0.031	<b>0.552</b>	<b>0.834</b>	<b>0.599</b>
LLaVA-v1.6-I	0.379	<b>0.396</b>	0.324	0.032	0.169	0.128	0.156
mPLUG-Owl2-I	0.257	0.257	0.169	0.083	0.078	0.164	0.120
InternLM-XC2-VL-I	0.348	0.376	<b>0.379</b>	<b>0.144</b>	0.034	0.108	0.123
GPT-4V-I	<b>0.470</b>	0.244	0.340	0.122	<b>0.531</b>	<b>0.761</b>	<b>0.714</b>
LLaVA-v1.6-C	0.332	0.355	0.257	0.109	0.124	0.065	0.174
mPLUG-Owl2-C	0.409	0.334	0.318	0.013	0.199	0.122	0.130
InternLM-XC2-VL-C	0.332	0.411	0.267	<b>0.131</b>	0.165	0.556	0.546
GPT-4V-C	<b>0.818</b>	<b>0.830</b>	<b>0.786</b>	0.124	<b>0.639</b>	<b>0.881</b>	<b>0.771</b>
Multiple-stimulus Method							
LLaVA-v1.6-S	0.349	0.351	0.315	<b>0.241</b>	0.169	0.221	0.210
mPLUG-Owl2-S	0.385	0.428	0.297	0.104	0.124	0.061	0.228
InternLM-XC2-VL-S	0.484	0.420	0.241	0.015	0.047	0.044	0.154
GPT-4V-S	<b>0.824</b>	<b>0.844</b>	<b>0.747</b>	0.037	<b>0.397</b>	<b>0.715</b>	<b>0.461</b>
LLaVA-v1.6-I	0.337	0.380	0.356	<b>0.203</b>	0.152	0.033	<b>0.241</b>
mPLUG-Owl2-I	0.268	0.268	0.377	0.067	<b>0.196</b>	0.142	0.121
InternLM-XC2-VL-I	0.489	0.235	0.212	0.046	0.038	0.102	0.114
GPT-4V-I	<b>0.585</b>	<b>0.496</b>	<b>0.389</b>	0.023	0.168	<b>0.416</b>	0.201
LLaVA-v1.6-C	0.292	0.424	0.288	0.043	0.227	0.111	0.122
mPLUG-Owl2-C	0.377	0.406	0.376	<b>0.126</b>	0.214	0.166	0.084
InternLM-XC2-VL-C	0.500	0.466	0.273	0.038	0.031	0.037	0.148
GPT-4V-C	<b>0.761</b>	<b>0.806</b>	<b>0.754</b>	0.036	<b>0.537</b>	<b>0.817</b>	<b>0.679</b>

- **Analysis of the Two IQA Scenarios**
  - No open-source MLLMs achieve satisfactory IQA results in the FR scenario regardless of the prompting system used
- **Analysis of Psychophysical Prompting Methods**
  - For the three open-source MLLMs, the single-stimulus method is the optimal choice
  - GPT-4V benefits from multiple-image analysis, and performs optimally under the double-stimulus chain-of-thought prompting
- **Analysis of NLP Prompting Methods**
  - In-context prompting is little likely to bring performance gains to the three open-source MLLMs
  - CoT prompting consistently enhances the performance of GPT-4V in conjunction with three psychophysical testing protocols and across nearly all visual attributes



# Findings

## Input Prompt



The first image is <ImageHere>. The second image is <ImageHere>. Please tell me the main object of two images, respectively.

## Answers

### LLaVA-v1.6 (Mistral-7B)

The main object in the first image is a cat.  
The main object in the second image is a cat.

### mPLUG-Owl2 (LLaMA2-7B)

In the first image, the main object is a cat with blue eyes sitting in a field of flowers. In the second image, the main object is also a cat with blue eyes, but it has a collar and is surrounded by lavender plants.

### InternLM-XComposer2-VL (InternLM2-7B)

In the first image, the main object is a white cat. In the second image, the main object is a red car.

### GPT-4V

The main object in the first image is a cat with blue eyes sitting in a garden setting. The second image's main object is a classic red convertible car, which appears to be parked on a road with the sunset in the background.

- Training instructions without multi-image VQA

**Fig. 5:** Behaviors of different MLLMs in recognizing objects from multiple images.

Image 0



Mos: 30.2

Image 1



Mos: 24.7

### GPT-4V (standard prompting)

Score: 0 ❌

### GPT-4V (chain-of-thought prompting)

**Description:** Both images exhibit motion blur, however, the first image preserves more details and some parts of the structure, such as buildings, are recognizable despite the blur. The color reproduction appears to be natural with ambient lighting reflecting true to the time of day. The second image has significant blur affecting the entire frame, making it difficult to discern any specific details or structure. The colors and textures are largely indistinguishable due to this heavy blur.

Score: 1 ✅

Score 0: The second image has better visual quality than the first image

Score 1: The first image has better visual quality than the second image

- Prompting with chain-of-thought method can provide accurate and reasonable results

**Fig. 6:** Comparison of the double-stimulus standard and chain-of-thought promptings for GPT-4V in the NR scenario.

# Further Testing on Difficult Data

**Table 2:** Comparison of MLLMs with optimally suited prompting systems against expert IQA systems in the FR scenario. \* indicates that the model has been trained on the dataset.

Method	FR-KADID		Aug-KADID		TQD		SPCD	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PSNR	0.479	0.675	0.381	0.644	0.345	0.522	<b>0.576</b>	<b>0.570</b>
SSIM [50]	0.553	0.694	0.405	0.633	0.510	0.618	0.229	0.246
FSIM [67]	<b>0.704</b>	<b>0.762</b>	0.400	0.560	0.332	0.408	0.205	0.206
LPIPS [68]	0.477	0.654	0.547	0.654	0.469	0.511	0.280	0.252
AHIQ [22]	0.512	0.583	0.512	0.688	0.467	0.608	0.240	0.269
DISTS [11]	0.647*	0.740*	<b>0.701</b>	<b>0.696</b>	<b>0.911</b>	<b>0.901</b>	0.454	0.422
LLaVA-v1.6 [29]	0.112	0.218	0.198	0.213	0.180	0.226	0.037	0.008
mPLUG-Owl2 [62]	0.248	0.435	0.358	0.484	0.228	0.335	0.102	0.108
InternLM-XC2-VL [13]	0.246	0.336	0.235	0.404	0.299	0.421	<b>0.171</b>	0.143
GPT-4V [59]	<b>0.669</b>	<b>0.795</b>	<b>0.708</b>	<b>0.800</b>	<b>0.786</b>	<b>0.857</b>	0.122	<b>0.234</b>

- They perform poorly on color difference discriminating
- Instruction tuning on low-level visual data can boost MLLM quality-aware capability

**Table 3:** Comparison of MLLMs with optimally suited prompting systems against expert IQA systems in the NR scenario.

Method	SPAQ		NR-KADID		AGIQA-3K	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE [34]	0.551	0.616	0.385	0.555	0.610	0.651
MUSIQ [20]	0.769	<b>0.817</b>	0.567	0.653	0.686	0.588
MANIQA [58]	0.546	0.564	0.428	0.387	0.521	0.599
LIQE [70]	<b>0.781*</b>	0.752*	<b>0.866*</b>	<b>0.930*</b>	<b>0.703</b>	<b>0.693</b>
LLaVA-v1.6 [29]	0.317	0.305	0.428	0.370	0.503	0.573
mPLUG-Owl2 [62]	0.270	0.198	0.128	0.187	0.168	0.201
InternLM-XC2-VL [13]	0.580	0.540	0.454	0.361	0.608	0.590
Q-Instruct [54]	0.799*	0.783*	<b>0.635</b>	<b>0.613</b>	<b>0.853*</b>	<b>0.821*</b>
GPT-4V [59]	<b>0.845</b>	<b>0.843</b>	0.513	0.453	0.783	0.746



# Take-home Messages

- Different MLLMs admit different prompting systems to work optimally
  - This suggests the need for a systematic **re-evaluation** of the recent progress
- There is still ample room in improving MLLMs (including GPT-4V) for IQA, especially for **fine-grained quality discrimination and multiple-image quality analysis**
- Directly fine-tuning open-source MLLMs on datasets with image quality descriptions may not be an effective strategy due to the risk of catastrophic forgetting
  - Active continual learning and/or parameter-efficient tuning of MLLMs to strike a good balance between the specificity (to IQA) and the generality of open-source MLLMs

## Input Prompt



Please describe this image.

## Answers

### LLaVA-v1.5 (Vicuna-v1.5-7B)

*The image shows a tiger in a natural setting, likely a zoo or wildlife park. The tiger is standing on its hind legs, with its front paws raised as if it's stretching or perhaps in a playful or curious posture. The tiger's fur is a mix of orange and black stripes, typical of the species, and it appears to be looking directly at the camera with a relaxed expression. The background is blurred, but it suggests a grassy area with trees, indicating a habitat designed to mimic the tiger's natural environment. The lighting in the image is bright, suggesting it might be a sunny day.*

### Q-Instruct (based on LLaVA-v1.5 (Vicuna-v1.5-7B))

*y area, and there is a tree in the background. The scene is set in a zoo, where the tiger is likely to be cared for and monitored by zookeepers.*

# Catastrophic forgetting



EUROPEAN CONFERENCE ON COMPUTER VISION

M I L A N O  
2 0 2 4

# Thanks for your listening

Tianhe Wu

2024.9.9

[wth22@mails.tsinghua.edu.cn](mailto:wth22@mails.tsinghua.edu.cn)



清華大學

Tsinghua University



香港城市大學  
City University of Hong Kong

oppo



THE HONG KONG  
POLYTECHNIC UNIVERSITY  
香港理工大學