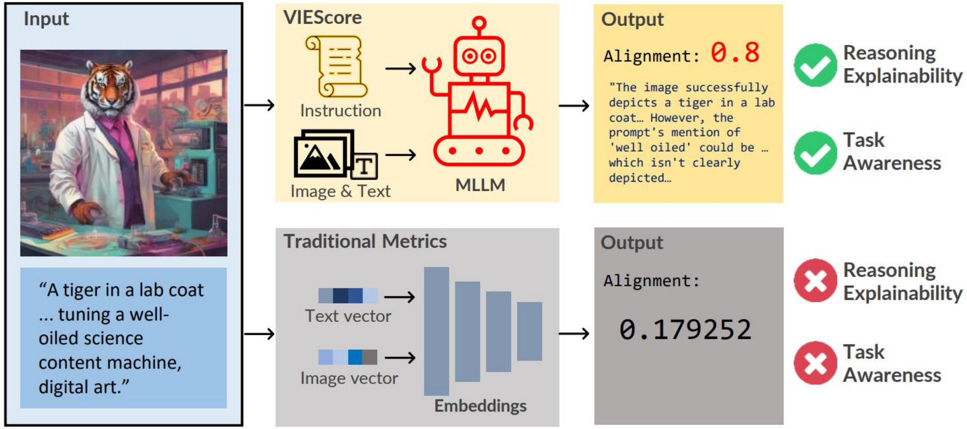


"A tiger in a lab coat ... tuning a well-oiled science content machine, digital art."

"The image successfully depicts a tiger in a lab coat… However, the prompt's mention of 'well oiled' could be … which isn't clearly depicted…

VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation

♠Max Ku, ♠Dongfu Jiang, ♠Cong Wei, ♥Xiang Yue, ♠Wenhu Chen

♠University of Waterloo, ♥IN.AI Research

tiger-ai-lab.github.io/VIEScore

Paper ID: #2258

ACL 2024
Bangkok, Thailand

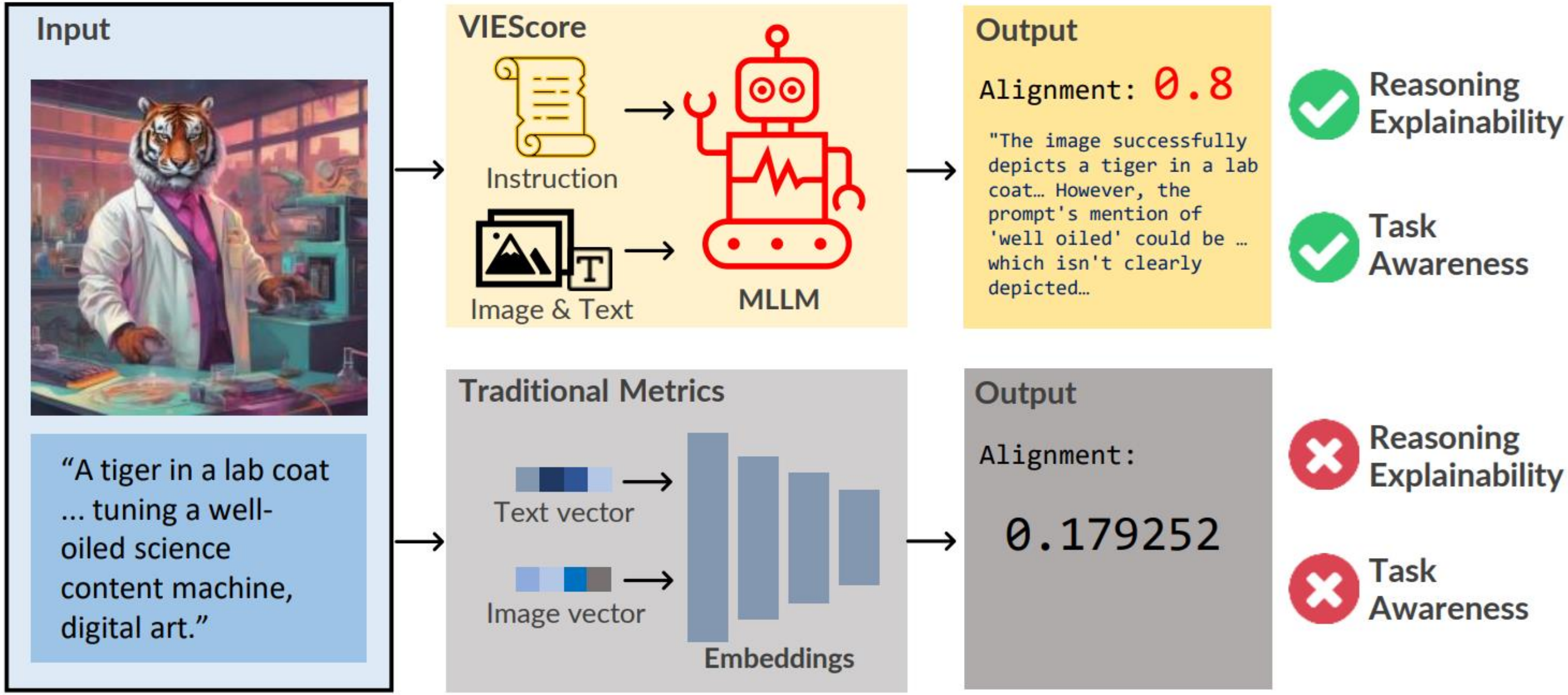
UNIVERSITY OF WATERLOO

Motivation

Metrics in the future would provide not just the score but also the rationale

- Understanding of judgment per instance is needed
- Traditional metrics are not task aware

Visual Instruction-guided Explainable Score (VIEScore)



Input

"A tiger in a lab coat ... tuning a well-oiled science content machine, digital art."

VIEScore

Instruction

Image & Text

MLLM

Output

Alignment: 0.8

"The image successfully depicts a tiger in a lab coat… However, the prompt's mention of 'well oiled' could be … which isn't clearly depicted…

✓ Reasoning Explainability

✓ Task Awareness

Traditional Metrics

Text vector

Image vector

Embeddings

Output

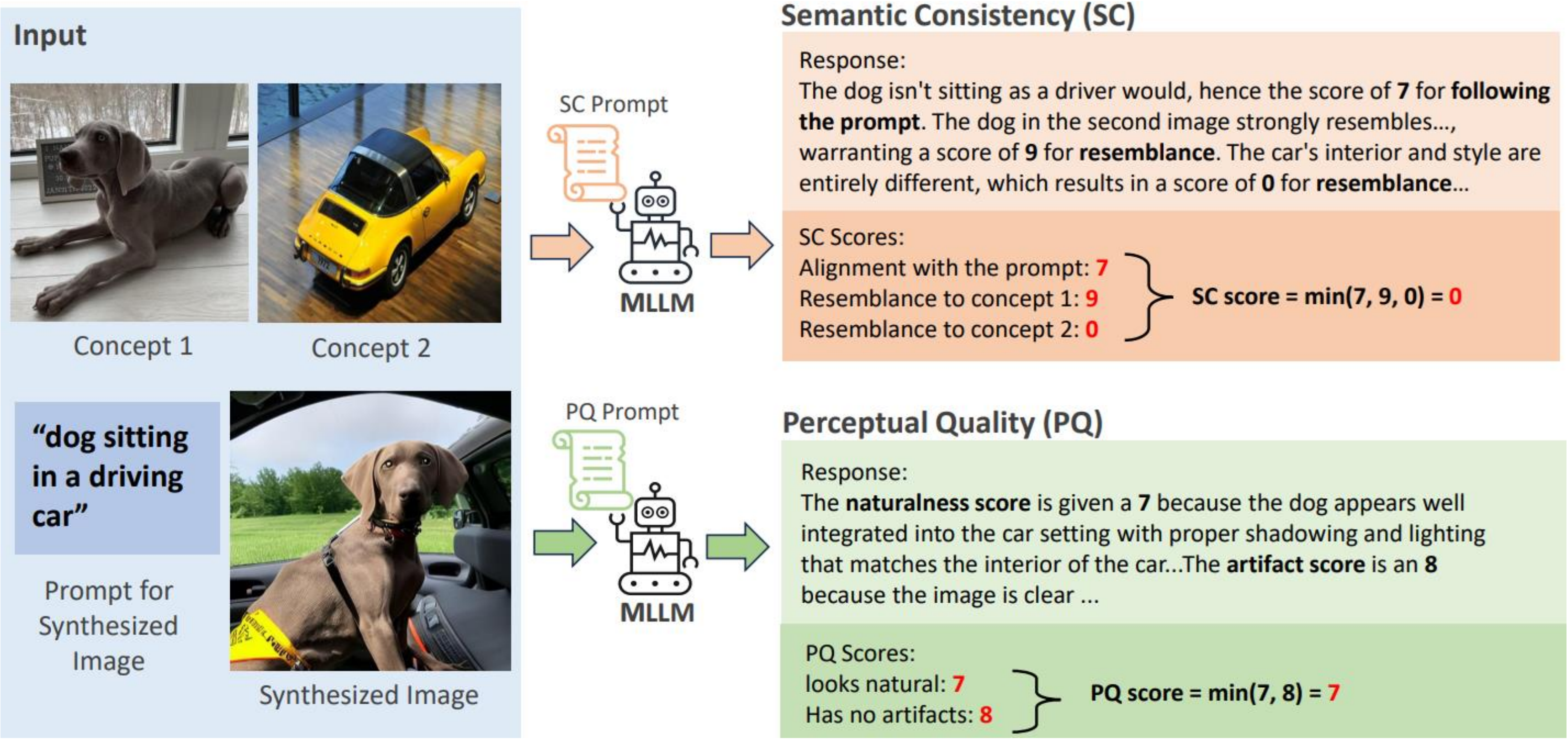
Alignment:

0.179252

✗ Reasoning Explainability

✗ Task Awareness

How VIEScore works?



Semantic Consistency (SC)

Response:
The dog isn't sitting as a driver would, hence the score of **7** for **following the prompt**. The dog in the second image strongly resembles..., warranting a score of **9** for **resemblance**. The car's interior and style are entirely different, which results in a score of **0** for **resemblance**...

SC Scores:
Alignment with the prompt: 7
Resemblance to concept 1: 9
Resemblance to concept 2: 0

SC score = min(7, 9, 0) = 0

Perceptual Quality (PQ)

Response:
The **naturalness score** is given a **7** because the dog appears well integrated into the car setting with proper shadowing and lighting that matches the interior of the car...The **artifact score** is an **8** because the image is clear ...

PQ Scores:
looks natural: 7
Has no artifacts: 8

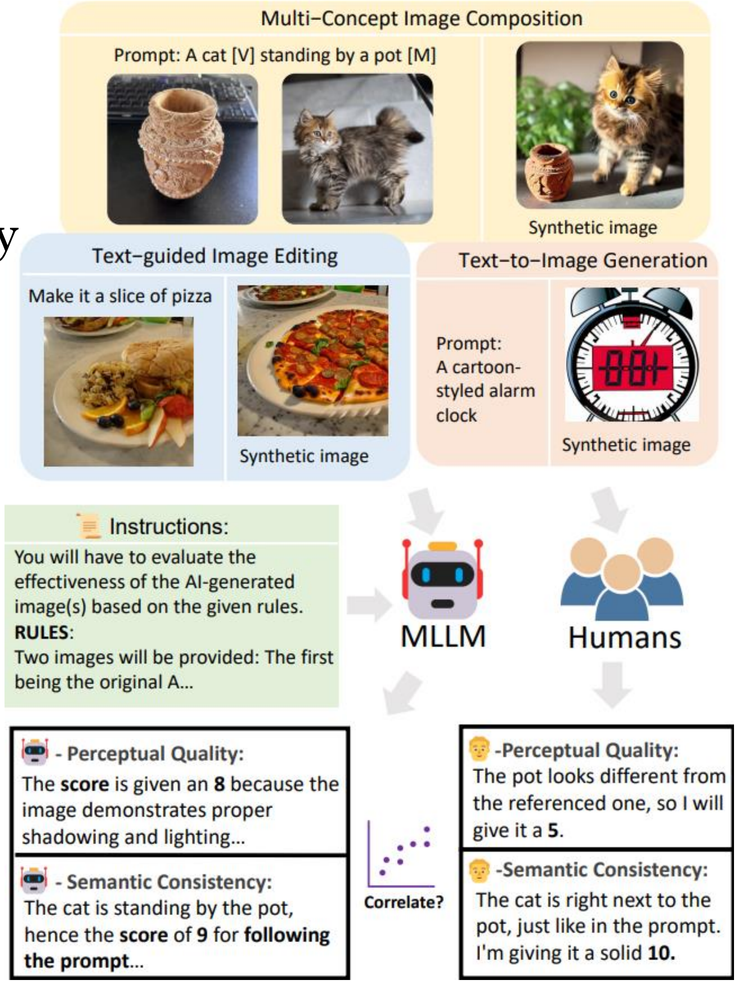
PQ score = min(7, 8) = 7

Experiment Setup

A wide range of image synthesis tasks study

- Correlation of VIEScore to Human
- V.S.
- Correlation of Traditional metrics to Human

Where can we get this kind of human annotation data?


Multi–Concept Image Composition
Prompt: A cat [V] standing by a pot [M]
Synthetic image
Text–guided Image Editing
Make it a slice of pizza
Synthetic image
Text–to–Image Generation
Prompt:
A cartoon-styled alarm clock
Synthetic image
MLLM
Humans
Correlate?

Instructions:
You will have to evaluate the effectiveness of the AI-generated image(s) based on the given rules.
RULES:
Two images will be provided: The first being the original A...

- Perceptual Quality:
The **score** is given an **8** because the image demonstrates proper shadowing and lighting...

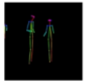
- Semantic Consistency:
The cat is standing by the pot, hence the **score** of **9** for **following the prompt**...

-Perceptual Quality:
The pot looks different from the referenced one, so I will give it a **5**.

-Semantic Consistency:
The cat is right next to the pot, just like in the prompt. I'm giving it a solid **10.**

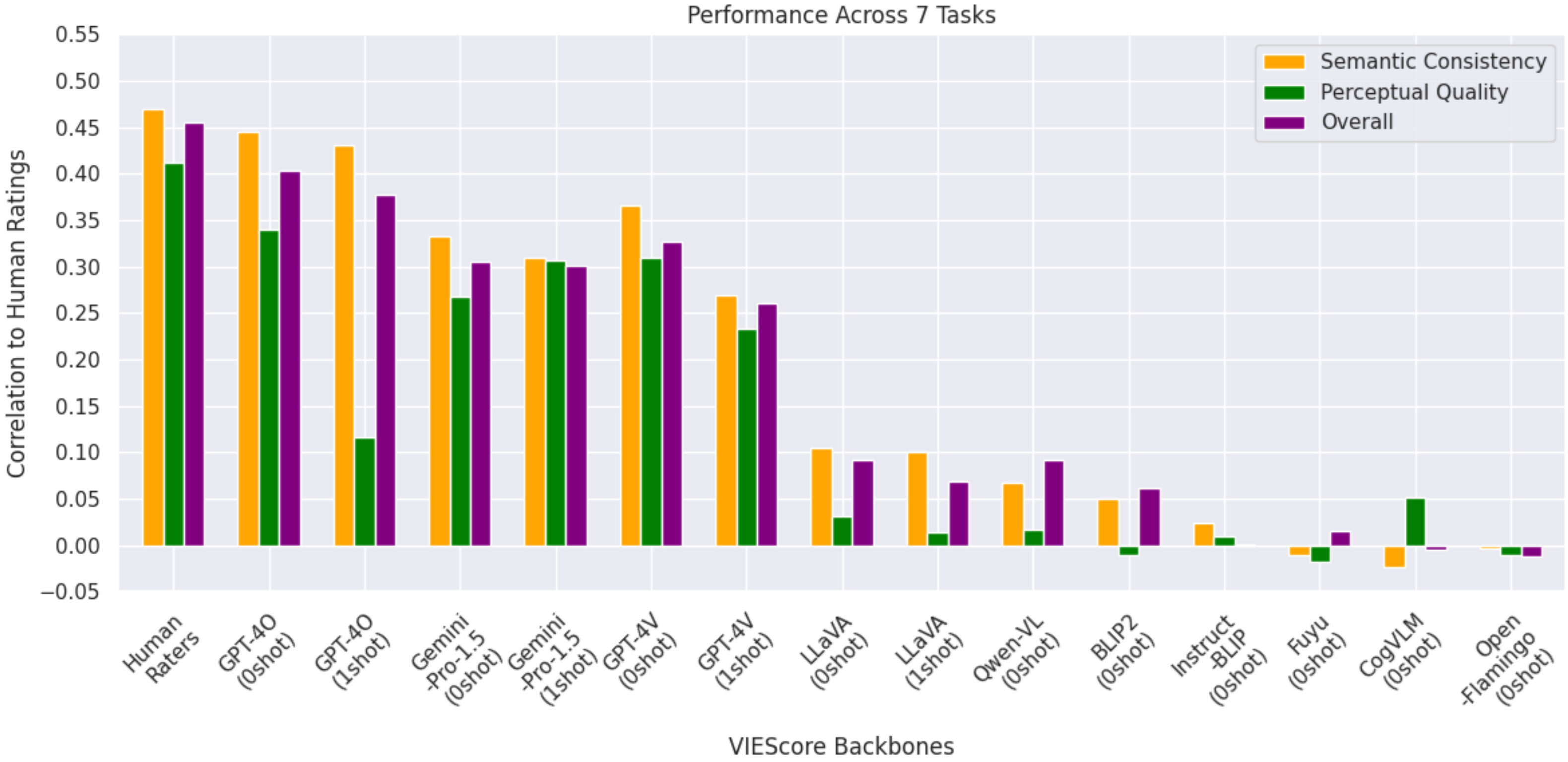
Experiment Setup (Cont.)

- Human data from ImagenHub(ICLR 2024)
- 29 Models across 7 tasks
  - Total 14403 annoations
- Each annotation 3 human metrics:
  - SC : Conditions-Image alignment
  - PQ: Realism and Natural sense
  - Overall: sqrt(SC x PQ)
- Human guideline is used as prompt

c_1	c_2	c_2	Task	y
A cartoon styled alarm clock	∅	∅	Text-to-Image Generation	
		Change frisbee to a football	Mask-guided Image Editing	
	Make it a slice of pizza instead of the sandwich	∅	Text-guided Image Editing	
	A [V] dog in the Versailles hall of mirrors	∅	Subject-Driven Image Generation	
		Replace glasses with [V] glasses	Subject-Driven Image Editing	
		A cat [V] standing by a pot [M]	Multi-Concept Image Composition	
	A small dog is curled up on top of the shoes	∅	Control-guided Image Generation	

ImagenHub: Standardizing the evaluation of conditional image generation models
https://arxiv.org/abs/2310.01596

Main Result

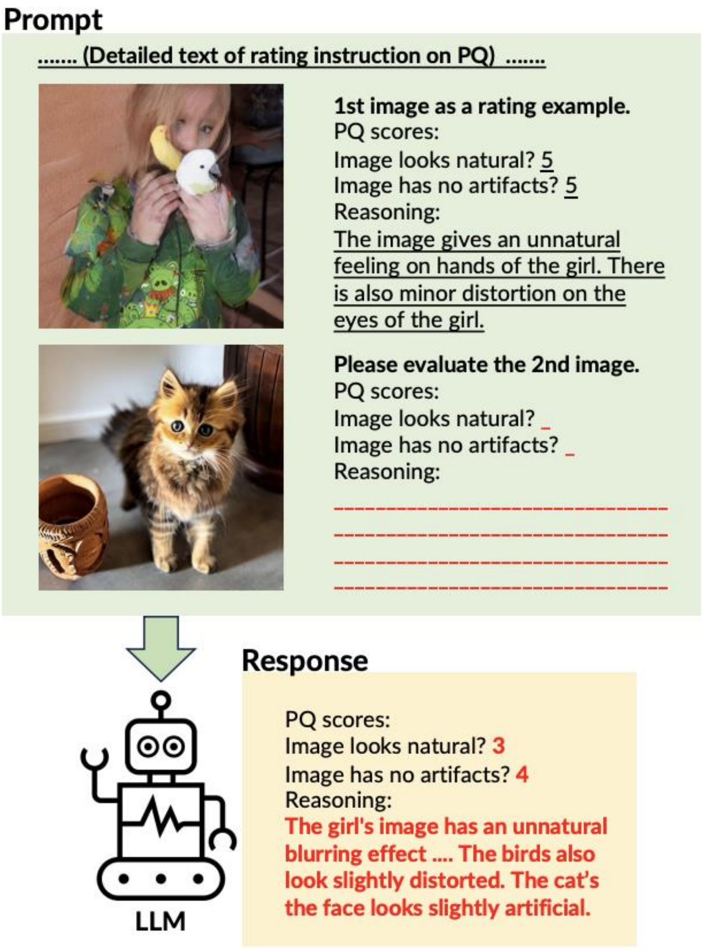


Performance Across 7 Tasks
Correlation to Human Ratings
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
Semantic Consistency
Perceptual Quality
Overall
Human Raters
GPT-4O (0shot)
GPT-4O (1shot)
Gemini -Pro-1.5 (0shot)
Gemini -Pro-1.5 (1shot)
GPT-4V (0shot)
GPT-4V (1shot)
LLaVA (0shot)
LLaVA (1shot)
Qwen-VL (0shot)
BLIP2 (0shot)
Instruct -BLIP (0shot)
Fuyu (0shot)
CogVLM (0shot)
Open -Flamingo (0shot)
VIEScore Backbones

Why one-shot setting achieve worse performance?

- MLLMs struggle in In-Context Learning when multiple images exists
  - Reasoning is affected

- Appears on all MLLMs we benchmarked



Prompt

.... (Detailed text of rating instruction on PQ) ....

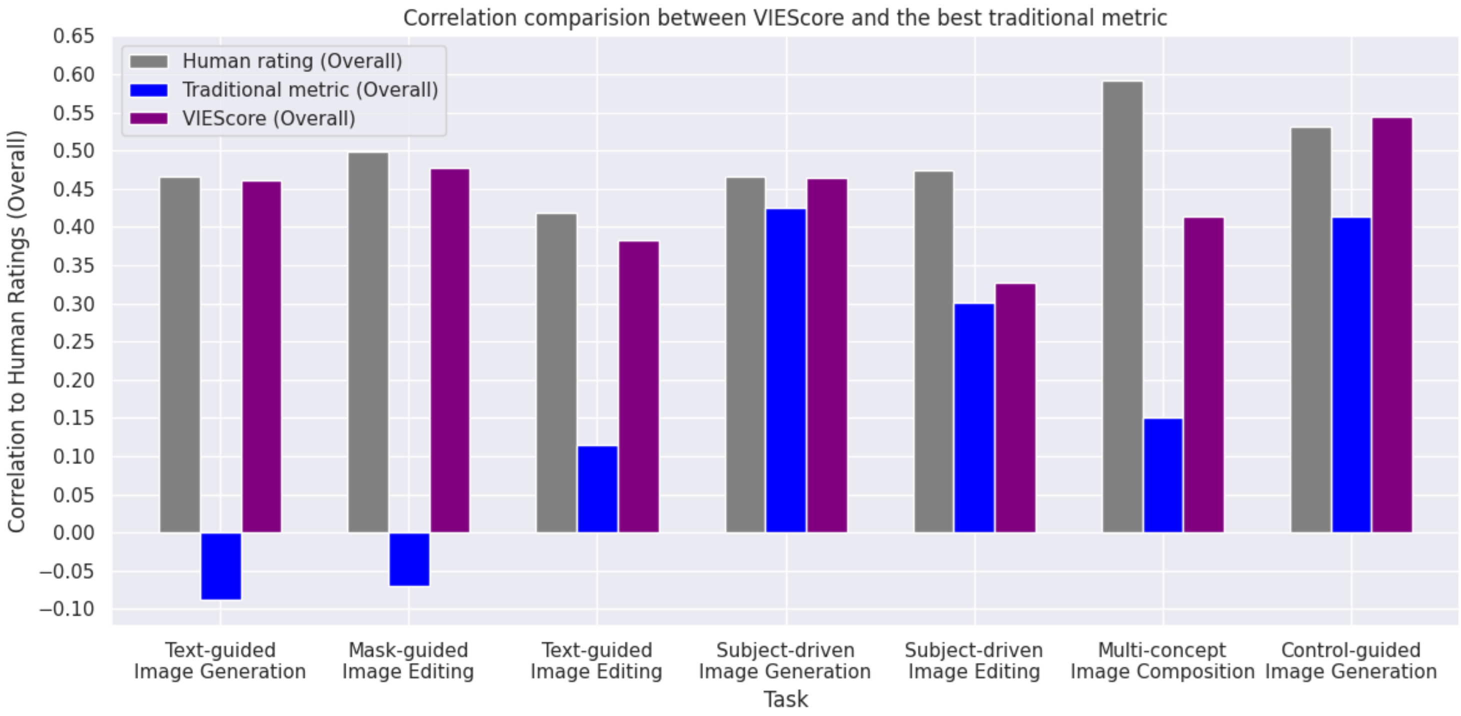
1st image as a rating example.
PQ scores:
Image looks natural? 5
Image has no artifacts? 5
Reasoning:
The image gives an unnatural feeling on hands of the girl. There is also minor distortion on the eyes of the girl.

Please evaluate the 2nd image.
PQ scores:
Image looks natural? _
Image has no artifacts? _
Reasoning:

Response

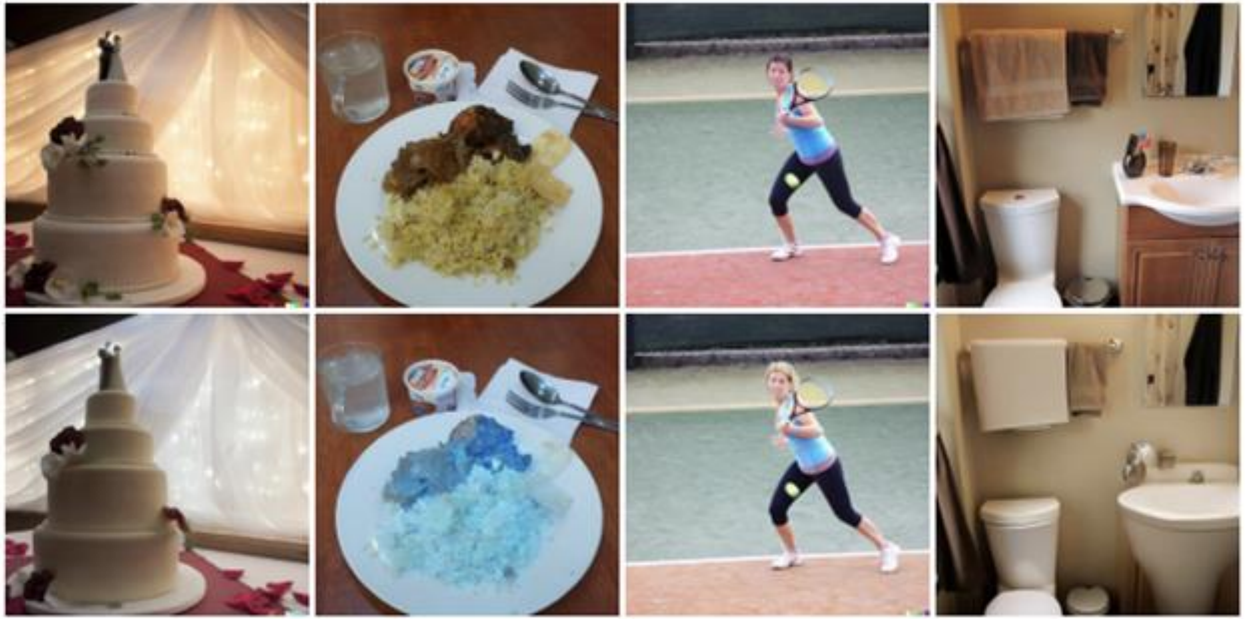
PQ scores:
Image looks natural? 3
Image has no artifacts? 4
Reasoning:
The girl's image has an unnatural blurring effect .... The birds also look slightly distorted. The cat's the face looks slightly artificial.

VIEScore achieve higher correlation than Traditional metrics



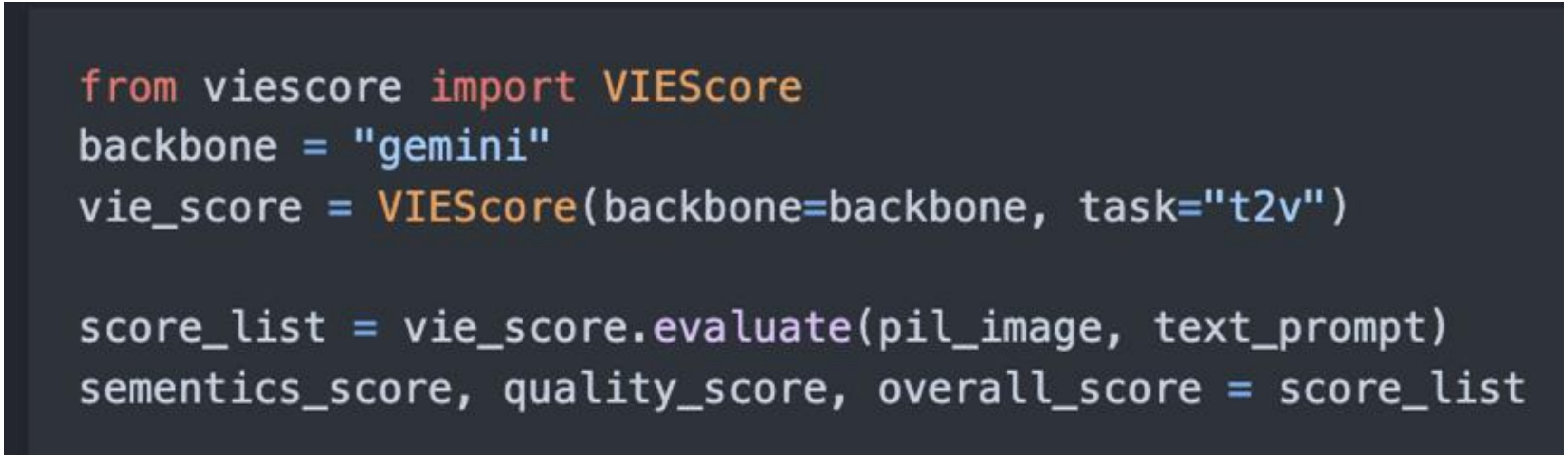
There’s still obstacles toward explainable metrics

- MLLMs struggles to see minor difference when given 2 similar images



Code Release

- Code available on GitHub
- Easy to use and modify



```
from viescore import VIEScore
backbone = "gemini"
vie_score = VIEScore(backbone=backbone, task="t2v")

score_list = vie_score.evaluate(pil_image, text_prompt)
sementics_score, quality_score, overall_score = score_list
```