VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation

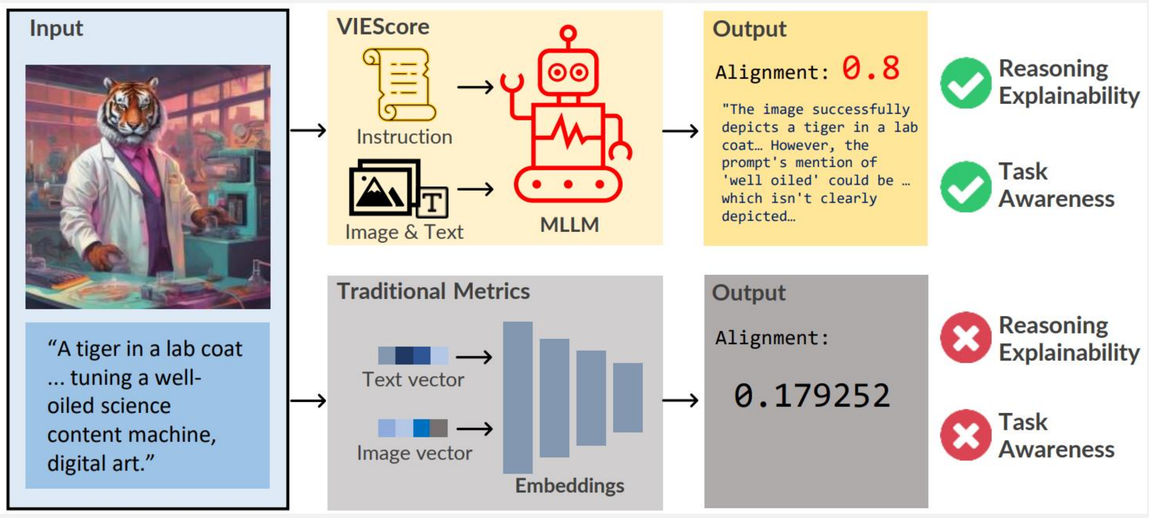
♠Max Ku, ♠Dongfu Jiang, ♠Cong Wei,
♥Xiang Yue, ♠Wenhu Chen
♠University of Waterloo, ♣IN.AI Research

ACL 2024 Bangkok, Thailand

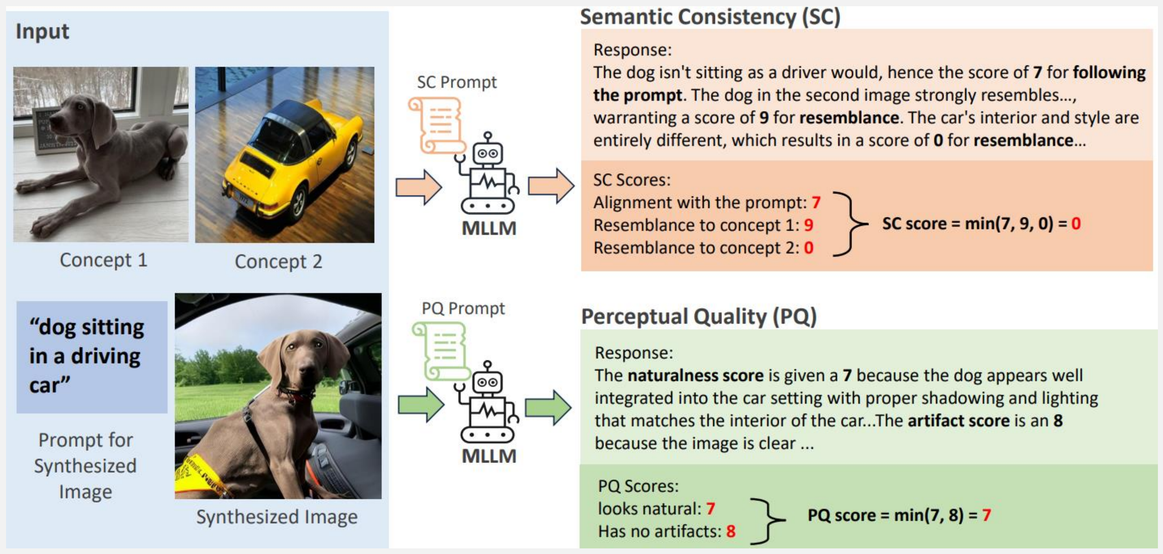
Code/Leaderboard

UNIVERSITY OF WATERLOO
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Visual Instruction-guided Explainable (VIE) Metric is needed for image synthesis tasks

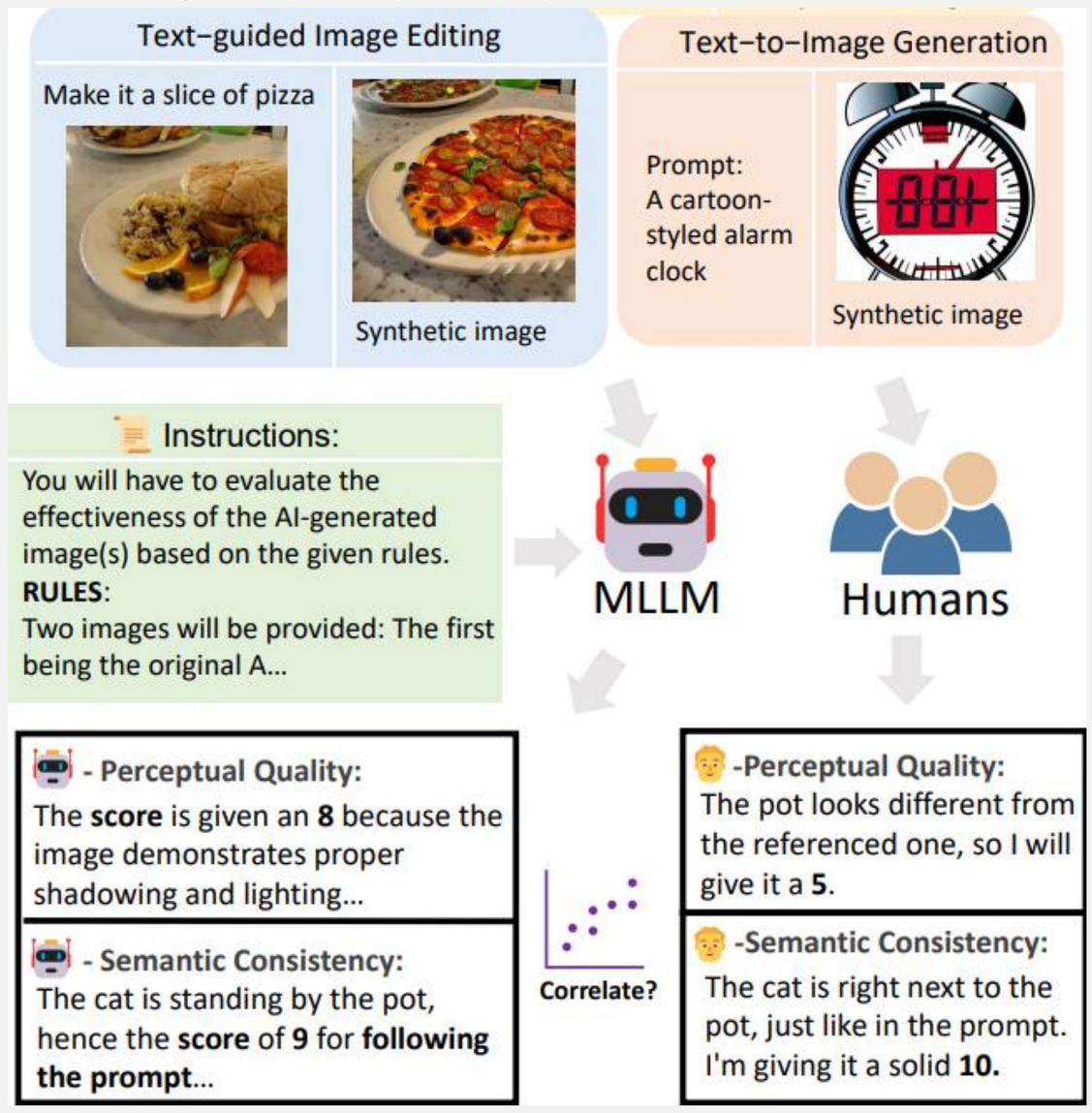


Method of VIEScore



Experiment Setup

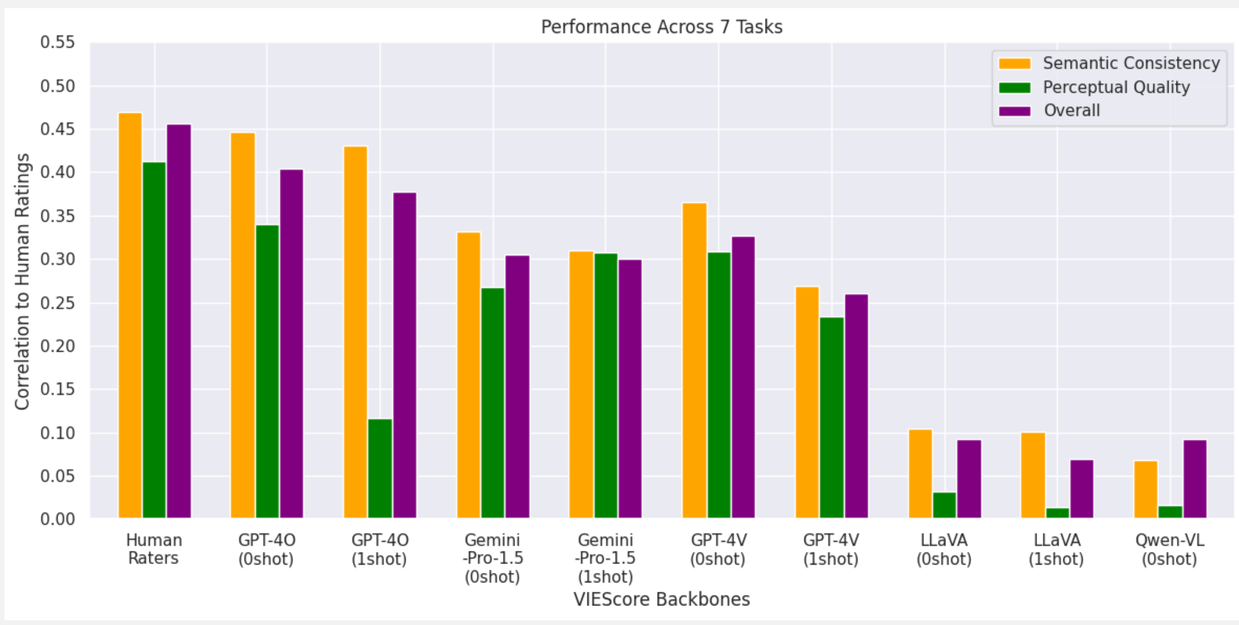
Apply VIEScore on the evaluation of a wide range of image generation and editing tasks (7 tasks) and compute the correlations with human



Human annotations from ImagenHub (ICLR2024), 14000+ instances of 29 models rating across 7 synthesis tasks

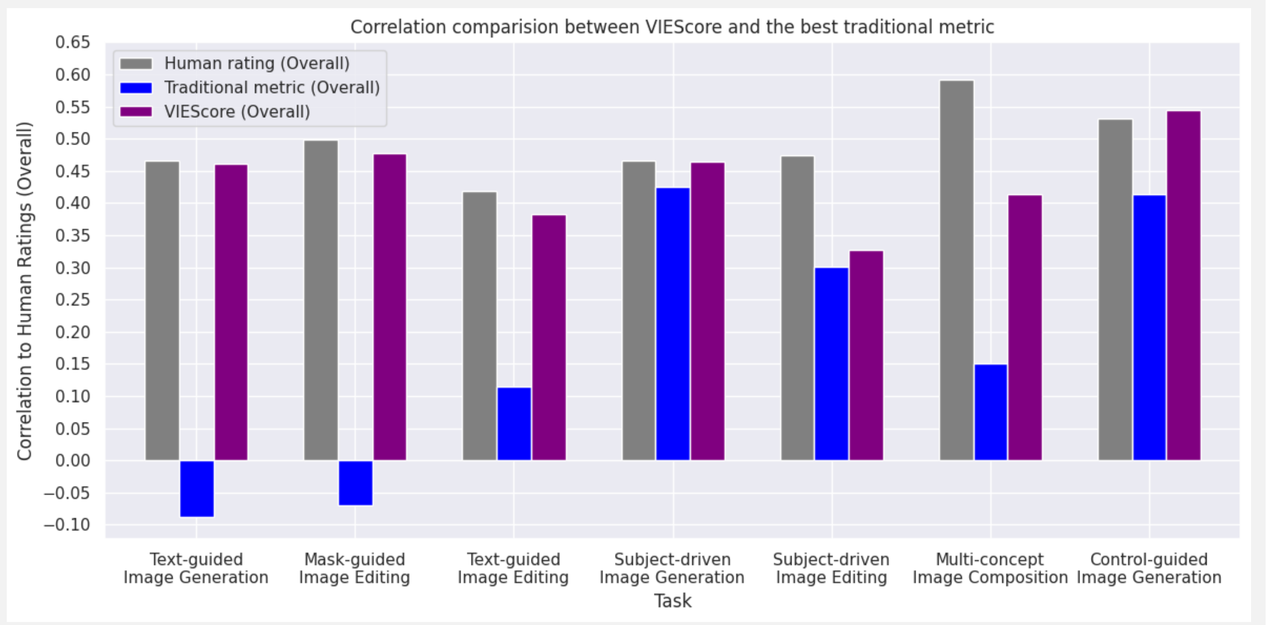
How effective is VIEScore with existing Multimodal LLMs?

VIEScore+GPT-4O achieve on par correlation with human ratings



VIEScore+Close-sourced MLLMs achieve better results in image evaluation.

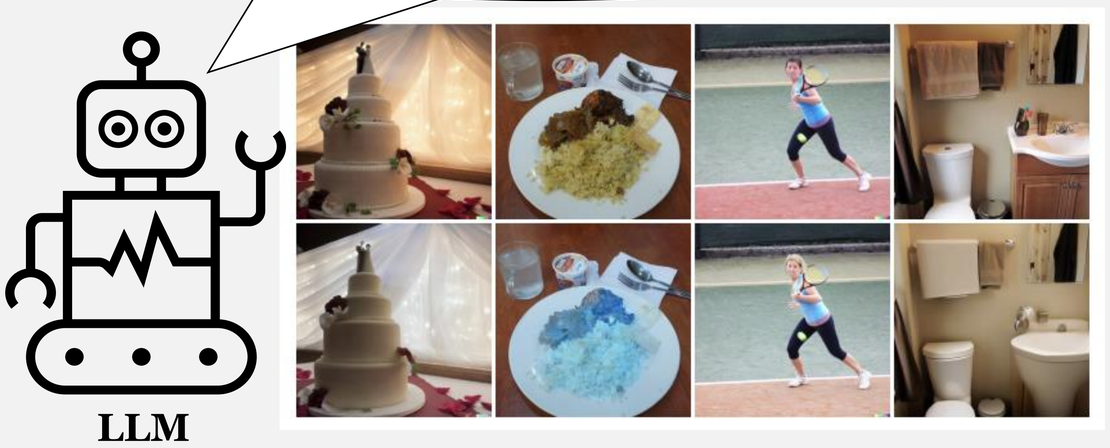
VIEScore excels in evaluating common image synthesis tasks



Traditional Metrics has poor correlation with human ratings on image.

Challenges and Obstacles Towards Explainable Metrics

General MLLMs struggle in image editing tasks



In-Context Learning with images confuses MLLMs