ImagenHub: Standardizing The Evaluation of Conditional Image Generation Models

♠Max Ku, ♠Tianle Li, †Kai Zhang, ♣Yujie Lu, ♥Xingyu Fu, ◊Wenwen Zhuang, ♠Wenhu Chen

♠University of Waterloo, †Ohio State University, ♣University of California Santa Barbara, ♥University of Pennsylvania, ◊Central South University

UNIVERSITY OF WATERLOO
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

To identify current progress in the field

Rapid research development of numerous image generation models.

However, significant inconsistencies in:

- Evaluation Datasets
- Inference methods
- Evaluation methodology

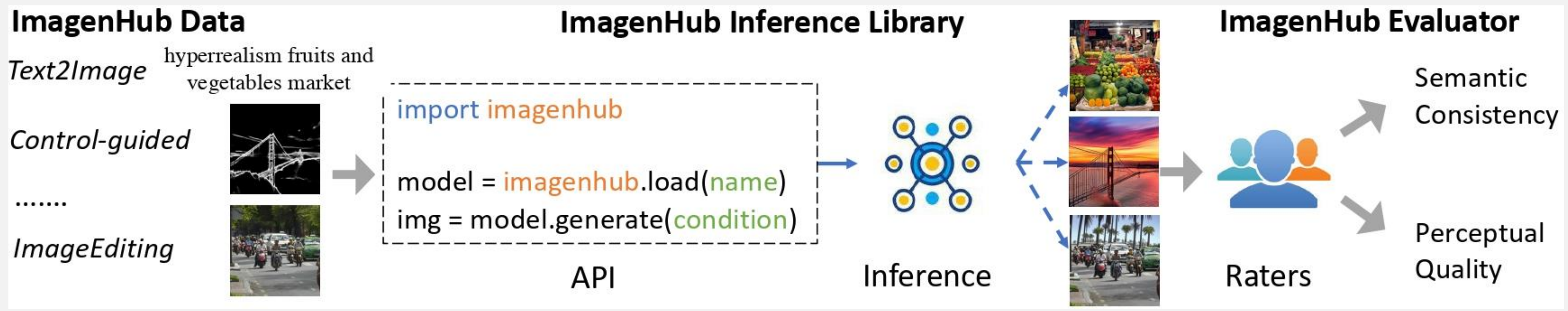
Leading difficulties in fairly comparing models

- Hinder the understanding of the true progress in the research field

Code, Visualization, etc.

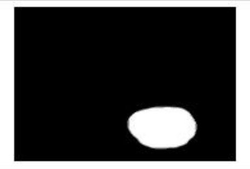
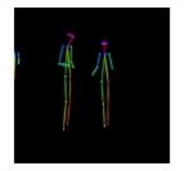
The evaluation platform we needed

ImagenHub is a python library to standardize the evaluation and analyze model reliability across seven common image generation tasks.



- Data: Curated high-quality human evaluation dataset for each task
- Library: Standardized inference pipeline for fair comparison
- Evaluator: Evaluated over 30 image models on two human metrics
  - Semantic Consistency (SC) : is image aligned with the condition(s)?
  - Perceptual Quality (PQ) : is image making sense and in good quality?

The Seven Tasks and Evaluation

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	

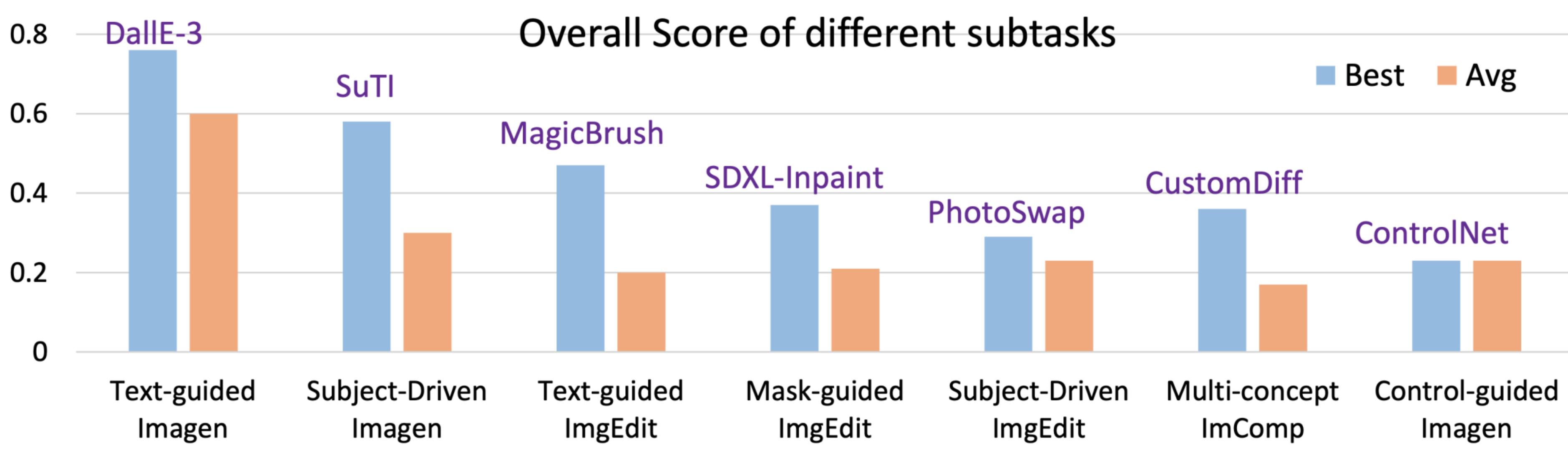
Three raters achieve high inter-worker agreement

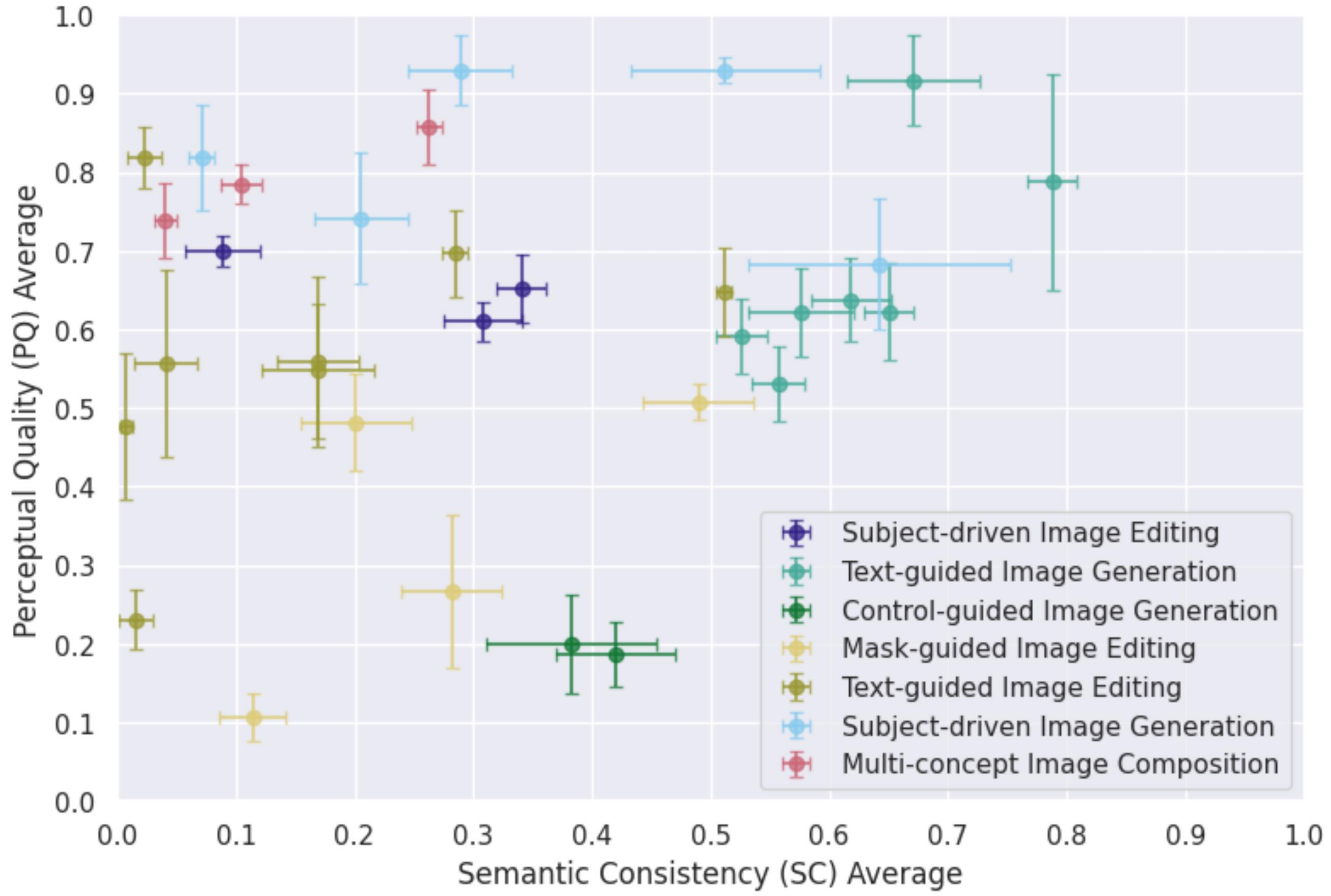
- Easy-to-follow guidelines and optimal options

Condition 1	Condition 2	Condition 3	SC rating
Inconsistent	Any	Any	0
Any	Inconsistent	Any	0
Any	Any	Inconsistent	0
Partially Consistent	Any	Any	0.5
Any	Partially Consistent	Any	0.5
Any	Any	Partially Consistent	0.5
Mostly Consistent	Mostly Consistent	Mostly Consistent	1.0

Subjects in image	Artifacts	Unusual sense	PQ rating
Unrecognizable	Any	Any	0
Any	Serious	Any	0
Recognizable	Moderate	Any	0.5
Recognizable	Any	Moderate	0.5
Recognizable	Little/None	Little/None	1.0

How robust are the image models?





Transparency: Results Hosted Online