Qi LIU, Shuanglin YANG, Zejian LI, Lefan HOU, Chenye MENG, Ying ZHANG, Lingyun SUN, 2025. Image generation evaluation: a comprehensive survey of human and automatic evaluations. *Frontiers of Information Technology & Electronic Engineering*, 26(7):1027-1065. https://doi.org/10.1631/FITEE.2400904

Image generation evaluation: a comprehensive survey of human and automatic evaluations

Key words: Image generation evaluation; Human evaluation; Automatic evaluation; Evaluation protocols; Evaluation aspects

Zejian LI
E-mail: zejianlee@zju.edu.cn
ORCID: https://orcid.org/0000-0001-5313-2742

Background

- Image generation models have made remarkable progress, and image evaluation is crucial for explaining and driving the development of these models.
- However, the evaluation methods have not been further developed in line with the advancement of image generation models, which is not conducive to their iterative improvement.

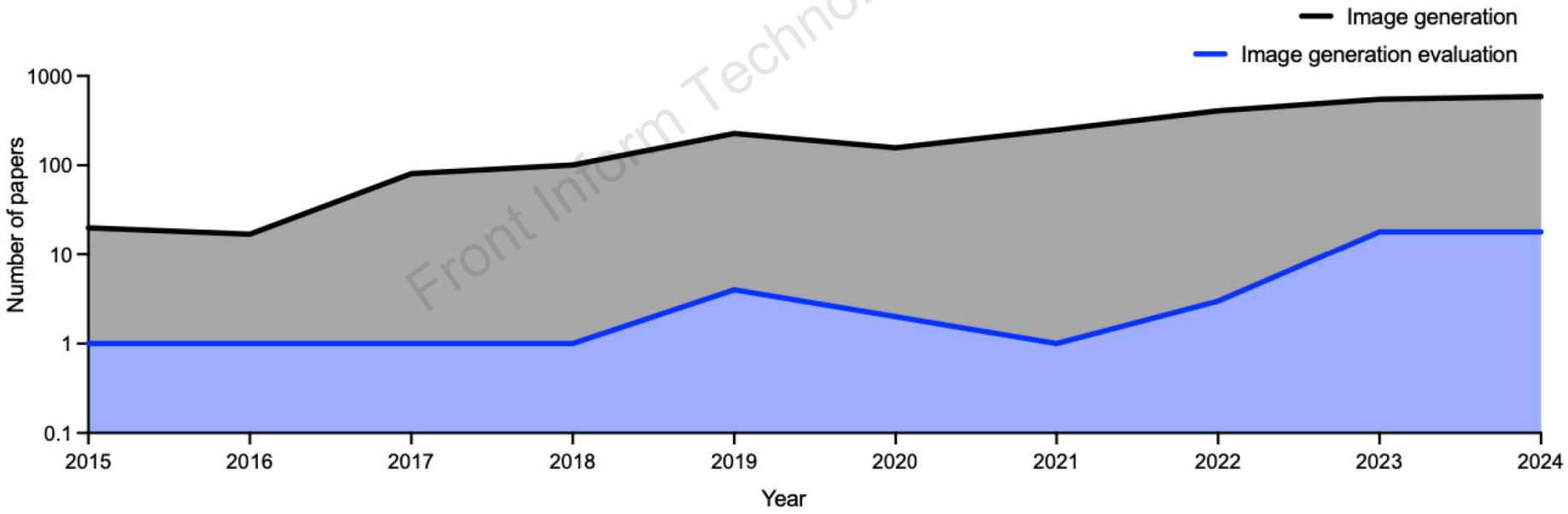
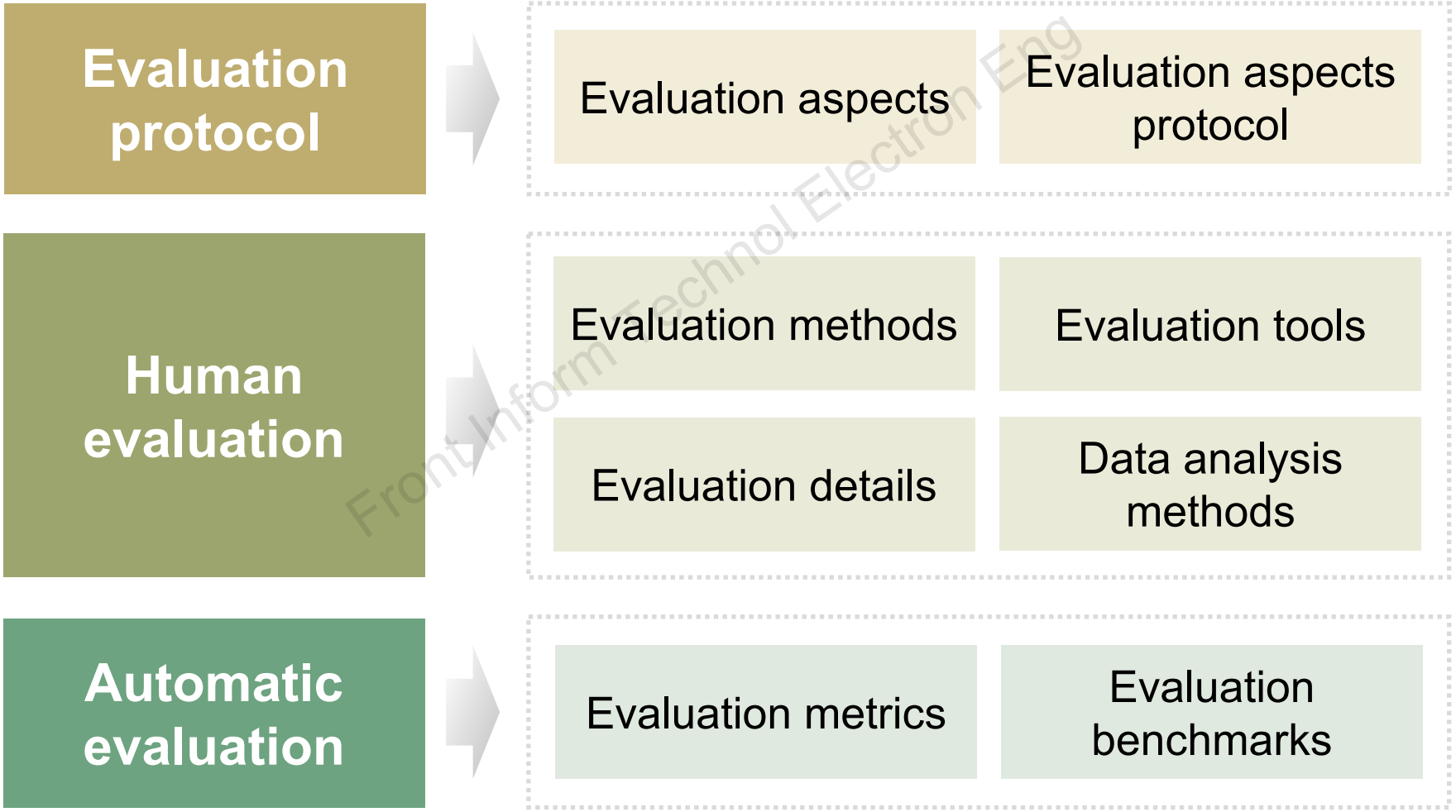


Fig. 1 Comparison of the number of published papers related to image generation and image generation evaluation in the last decade

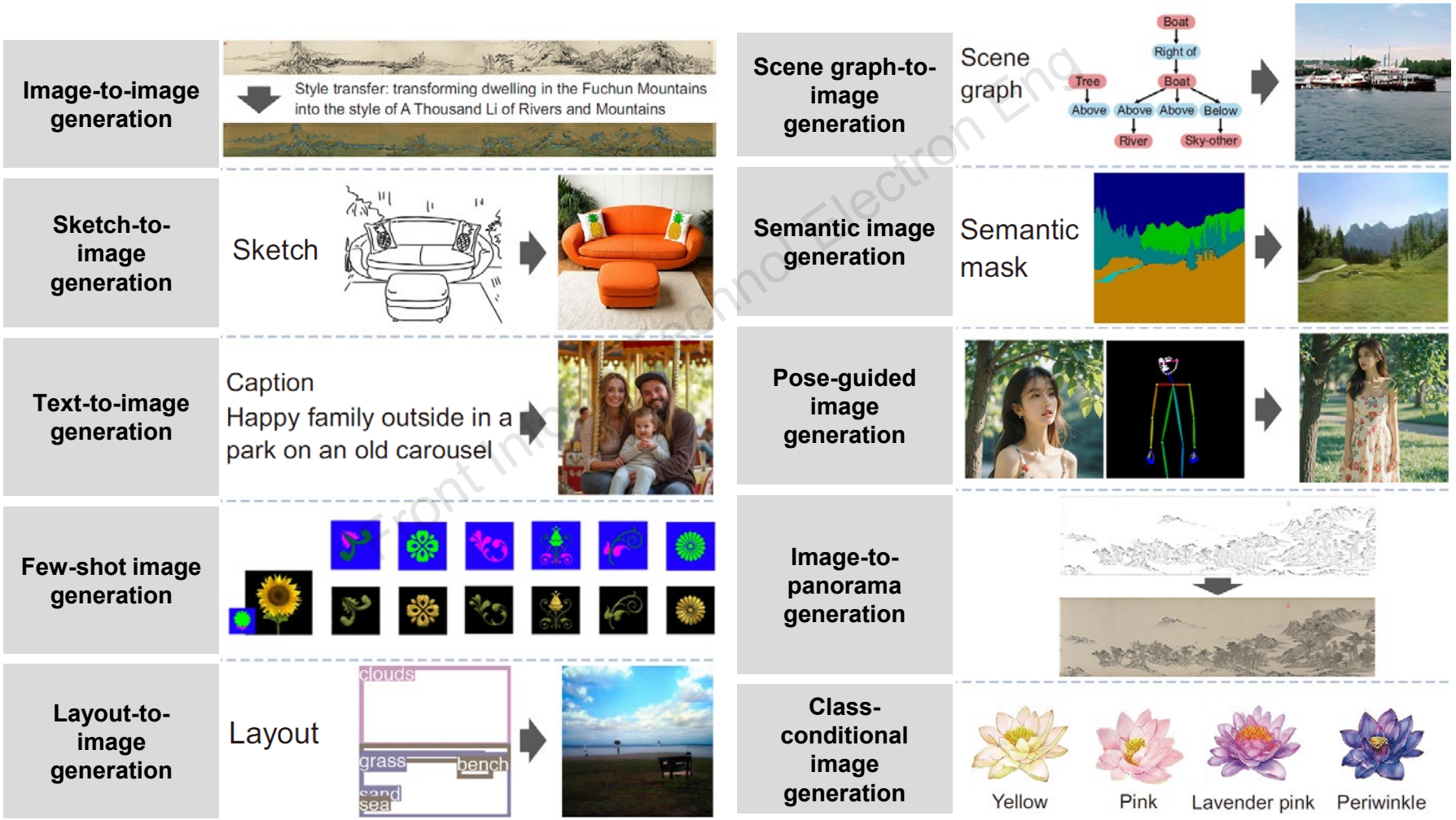
Objective

- This review aims to provide a survey on image generation evaluation, contributing to a systematic understanding of the field.



Various image generation tasks

- Ten categories of image generation tasks are defined based on different types of input conditions.



Protocol for evaluation aspects

- The evaluation metrics vary across different image generation tasks. We extract six commonly important evaluation aspects.

Fidelity	Fidelity of an image
Consistency	Representing how well the generated image matches the input content (such as text and sketch)
Recognizability	Degree to which the objects described in the input content can be accurately identified in the image
Diversity	Differences and richness of changes between images
Overall quality	Comprehensive assessment of an image as a whole
User preference	User's preference and inclination towards images

Protocol for evaluation aspects

- We summarize the automatic and human evaluation metrics used for each image generation task, and propose a protocol including the subjective and objective evaluation aspects.

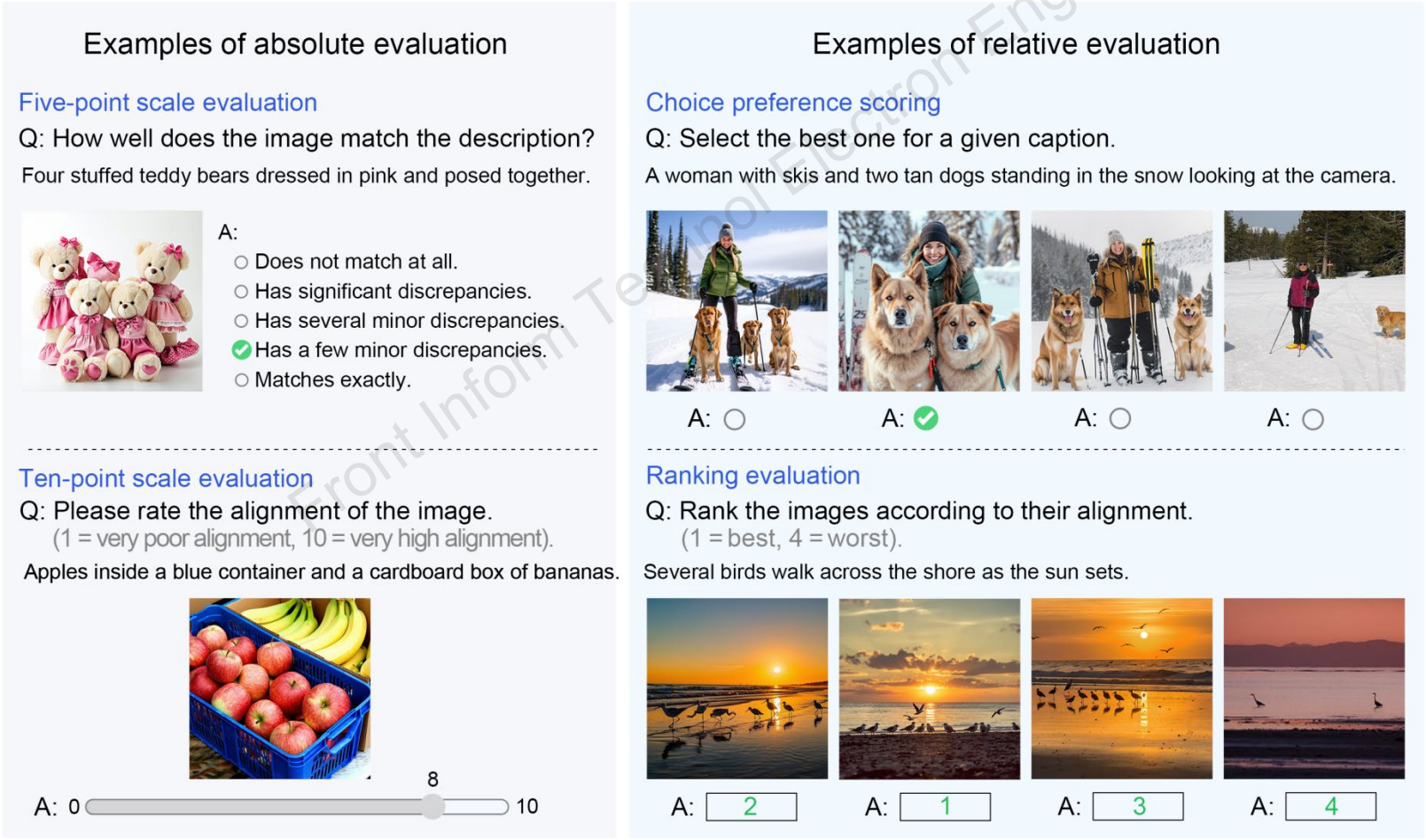
Table 3 Protocol for the subjective and objective evaluation aspects of image generation

Task	Fidelity	Consistency	Recognizability	Overall quality	User preference	Diversity
Image-to-image generation	✓△	✓△	△	✓△	✓	
Image super-resolution			△	✓△		
Image inpainting				✓△		
Style transfer	✓	✓△		✓△	✓	
Image-to-cartoon generation		△			✓	
Image colorization	✓△					
Attribute manipulation	✓	✓△		✓△		
Semantic manipulation	✓△	△		✓△		
Image dehazing	✓△		△	✓△		
Image deblurring				△		
Low-light enhancement				△		
Sketch-to-image generation	✓△	✓△		✓△		△
Text-to-image generation	✓△	✓△	✓△	✓△	✓△	✓△
Few-shot image generation	✓△			✓△		✓△
Layout-to-image generation	✓△	✓△	△	✓△	✓	△
Scene graph-to-image generation	△	✓△	✓△	△	✓	✓△
Semantic image generation	✓△	✓△		△	✓	✓△
Pose-guided image generation	✓△	△		✓△		△
Image-to-panorama generation	△	△		✓△	✓	△
Class-conditional image generation	✓△		△	△		△

✓and △ denote the evaluation aspects in human evaluation and automatic evaluation, respectively.

Human evaluation

- We present the first comprehensive analysis of human evaluation in image generation, covering evaluation methods, tools, details, and data analysis methods.



Automatic evaluation

- We systematically review automatic evaluation metrics and benchmarks in the field of image generation, with a particular focus on the progress made over the past five years.

Table 4 Comparison of evaluation benchmarks (focusing on text-to-image)

Benchmark	Year	SS	Metric	Human	Auto	PN	SHLC	NSTC
PaintSkills	2022	5	3	Yes	Yes	7330	No	–
DrawBench	2022	2	0	Yes	No	200	No	11
PartiPrompts	2022	2	0	Yes	No	1600	Yes	11
EntityDrawBench	2022	2	0	Yes	No	250	No	–
Multitask	2022	3	0	Yes	No	90	Yes	32
TISE	2022	3	5	No	Yes	–	No	–
HRS-Bench	2023	13	17	Yes	Yes	45 000	Yes	–
TIFA	2023	1	0	Yes	No	4081	No	12
HEIM	2023	12	25	Yes	Yes	~500 000	–	–
T2I-CompBench	2023	6	5	Yes	Yes	6000	Yes	3
GenAI-bench	2024	8	0	Yes	No	1600	Yes	–
PhyBench	2024	4	0	Yes	Yes	700	Yes	24

In this table, specified skill (SS) refers to the number of abilities that the benchmark is designed to evaluate the model. For example, PaintSkills evaluates five abilities: object recognition, object counting, spatial relationship understanding, gender bias, and skin tone bias. Although TIFA-bench identified deficiencies in models' counting and spatial relationship abilities, its design and evaluation only focus on model fidelity. Metric refers only to automatic evaluation indicators and not to quantitative human evaluations. PN refers to prompt number. Specified hardness level or challenge (SHLC) indicates whether the prompts are divided into corresponding difficulty levels or specific challenges during the design. NSTC refers to the number of specified tasks or challenges to which these prompts correspond, such as tasks generating a specified number of objects or handling absurd requests

Challenges

➢ Evaluation protocols

Evaluation aspects:

- Lack of generally accepted guidelines for selecting evaluation aspects
- Lack of comprehensive and universal evaluation benchmarks

Datasets:

- Accessibility threshold
- Long-tail effect
- Lack of high-quality, large-scale, and open datasets

➢ Evaluation methods

Human evaluation:

- Individual differences among participants
- Temporal and financial cost
- Lack of empirical research on evaluation methods

Automatic evaluation:

- Scope of applicability
- Reliability of evaluation results
- Interpretability of evaluation results

Future directions

➢ Evaluation protocols

Evaluation aspects:

- Developing widely accepted guidelines for selecting evaluation aspects
- Multidimensional evaluation
- Collecting diverse types of information

Datasets:

- Manual methods
- Automatic annotation methods
- Direct data generation methods

➢ Evaluation methods

Human evaluation:

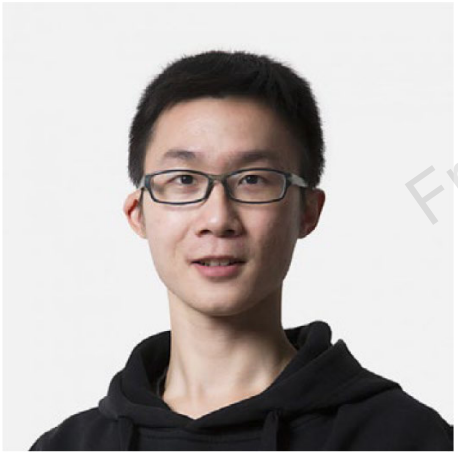
- Developing detailed evaluation instructions
- Standardized protocols
- Reliable quantitative tools
- Annotation quality management
- …

Automatic evaluation:

- Using multiple metrics
- Collecting large amounts of fine-grained data and diverse data types
- …

Author Bio

Qi Liu is a Ph.D. student in Artificial Intelligence at Zhejiang University. Her research interests include image generation evaluation and human−computer interaction.



Zejian Li is a Platform Top 100 researcher at the School of Software Technology, Zhejiang University, and serves as a Ph.D. supervisor. His research interests include training and distillation of diffusion models, interactive human−AI co-creation, and evaluation of generative content. His work has been published in top-tier international conferences such as ICLR, CVPR, ICCV, and AAAI.