Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

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

Qi LIU[1], Shuanglin YANG[2], Zejian LI[†‡1], Lefan HOU[3], Chenye MENG[3], Ying ZHANG[1], Lingyun SUN[3]

[1]*School of Software Technology, Zhejiang University, Ningbo 315100, China*
[2]*School of Computer Science and Engineering, Southeast University, Nanjing 211189, China*
[3]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*
[†]E-mail: zejianlee@zju.edu.cn

Received Oct. 12, 2024; Revision accepted Jan. 24, 2025; Crosschecked Apr. 25, 2025

Abstract: Image generation models have made remarkable progress, and image evaluation is crucial for explaining and driving the development of these models. Previous studies have extensively explored human and automatic evaluations of image generation. Herein, these studies are comprehensively surveyed, specifically for two main parts: evaluation protocols and evaluation methods. First, 10 image generation tasks are summarized with focus on their differences in evaluation aspects. Based on this, a novel protocol is proposed to cover human and automatic evaluation aspects required for various image generation tasks. Second, the review of automatic evaluation methods in the past five years is highlighted. To our knowledge, this paper presents the first comprehensive summary of human evaluation, encompassing evaluation methods, tools, details, and data analysis methods. Finally, the challenges and potential directions for image generation evaluation are discussed. We hope that this survey will help researchers develop a systematic understanding of image generation evaluation, stay updated with the latest advancements in the field, and encourage further research.

Key words: Image generation evaluation; Human evaluation; Automatic evaluation; Evaluation protocols; Evaluation aspects
https://doi.org/10.1631/FITEE.2400904 **CLC number:** TP391.4

1 Introduction

Image generation models have undergone significant advancement; therefore, the methods used for their evaluation must be continuously updated to ensure reliable results. The advancement of deep learning has facilitated aligning images and generating them from various data types such as texts, sketches, scene graphs, and layout graphs (Elasri et al., 2022). These images have been used in various fields, including medicine (Elasri et al., 2022), fashion (Elasri et al., 2022), material design (Yang L et al., 2023), media (Wu JY et al., 2023), and e-commerce (Wu JY et al., 2023). Quality must be ensured as it directly impacts the recipient's visual experience (Zhao K et al., 2023). The performance of image generation models must be reliably evaluated for their further development (Ioannou and Maddock, 2024; Jayasumana et al., 2024). 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 compares the number of published papers related to image generation and image generation evaluation in the last decade. Moreover,

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (No. 62006208) and the Provincial Key Research and Development Plan of Zhejiang Province (No. 2024C01250(SD2))
ORCID: Qi LIU, https://orcid.org/0000-0001-8784-7404; Zejian LI, https://orcid.org/0000-0001-5313-2742

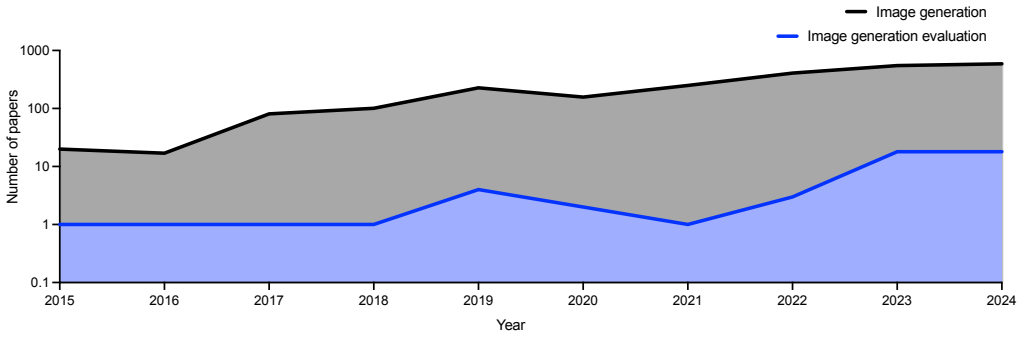

Fig. 1 Comparison of the number of published papers related to image generation and image generation evaluation in the last decade. Results showed that research on image generation evaluation has not kept pace with that on image generation. By checking the abstracts and key words, we retained 2399 papers on image generation and 46 papers on image generation evaluation. All papers were published in top-tier conferences relevant to computer vision, including the Conference on Computer Vision and Pattern Recognition (CVPR), Neural Information Processing Systems (NeurIPS), the International Conference on Learning Representations (ICLR), the International Conference on Computer Vision (ICCV), and the International Conference on Machine Learning (ICML), updated as of Dec. 25, 2024

advanced image generation techniques and the use of diverse tasks have saturated the existing metrics, thereby creating new evaluation demands. Therefore, image generation evaluation has been extensively studied (Zhao K et al., 2023). This paper summarizes such studies and presents the latest developments in this field. We aim to assist researchers in systematically understanding the progress in image generation evaluation and identifying promising research directions.

Image generation tasks are diverse, with varying evaluation aspects. In text-to-image generation, visual fidelity and semantic alignment are crucial evaluation aspects (Frolov et al., 2021; Otani et al., 2023), because understanding complex text and accurately translating it into visual are challenging. In contrast, fidelity, recognizability, and diversity are important factors in layout-to-image generation, and generating diverse images featuring multiple complex objects is a major challenge (Zhao B et al., 2019). Herein, the differences between various image generation tasks are explored and their distinct evaluation aspects are highlighted.

Both human and automatic evaluations are central to image evaluation (Zhu WH et al., 2018; Zhai and Min, 2020; Ma and Fang, 2021; Wang ZH et al., 2021; Gao YX et al., 2022); however, most studies have preferred using automatic evaluation over human evaluation owing to its high efficiency, objectivity, reproducibility, and cost-effectiveness (Zhou S et al., 2019; Frolov et al., 2021; Ioannou and Maddock, 2024; Ku et al., 2024). Although human evaluation is crucial and equally necessary, it has not been extensively studied and only a limited number of methods have been reported. Human evaluation is crucial and equally necessary. Automatic evaluation often fails to accurately capture all human perceptions, yielding inconsistent results compared with human evaluation (Wang ZH et al., 2021; Wang JR et al., 2023). In contrast, the success of an image generation model is ultimately determined via human evaluation as the generated images are evaluated by humans (Xu QQ et al., 2012; Zhu WH et al., 2018; Zhai and Min, 2020; Ding et al., 2021). To this end, this paper provides a systematic review of human evaluation in image generation tasks.

Although image generation evaluation has been extensively studied (Zhai and Min, 2020; Frolov et al., 2021; Wang ZH et al., 2021; Ioannou and Maddock, 2024), comprehensive and up-to-date reviews covering the evaluation of various image generation tasks are still lacking. Existing studies can be categorized into three types: (1) Surveys restricted to automatic evaluation offer limited insight into human evaluation. For instance, Zhai and Min (2020) conducted a survey focusing on automatic image quality assessment rather than comprehensive image generation evaluation, offering limited discussion on human evaluation and user studies. (2) Task-specific surveys provide insights into image evaluation but have limited applicability. For instance, Ioannou and Maddock (2024) discussed methods and metrics for image

evaluation, including side-by-side comparisons, user studies, and automatic metrics, but focused solely on neural style transfer. (3) Broader surveys often provide brief overviews of image generation evaluation, which are insufficient. For instance, Wang ZH et al. (2021) reported performance evaluation in terms of image super-resolution, covering benchmark datasets, performance metrics, and operational channels. Frolov et al. (2021) reviewed well-known metrics for quality evaluation and text alignment in text-to-image generation, along with user studies and challenges faced by these evaluation methods.

This review aims to provide a survey on image generation evaluation, contributing to a systematic understanding of the field. The main contributions are as follows:

1. Image generation evaluation is comprehensively surveyed, and 10 image generation tasks based on input conditions are summarized (Wang L et al., 2020; Alqahtani et al., 2021; Elasri et al., 2022; Pang et al., 2022; Croitoru et al., 2023). Both automatic and human evaluation methods are reviewed, providing the first in-depth analysis of human evaluation in image generation.

2. A new protocol for subjective and objective evaluation aspects is proposed, identifying six common important aspects: fidelity, consistency, recognizability, overall quality, user preference, and diversity. Additionally, automatic and human evaluation criteria are aligned with specific image generation types.

3. A taxonomy is developed, and the metrics and benchmarks for automatic evaluation of various image generation tasks are comparatively analyzed. Alongside classic evaluation metrics, the influential evaluation aspects from the past five years are highlighted.

4. Current challenges and potential research directions for image generation evaluation are discussed, including evaluation protocols and methods.

We focused on the papers published in the past decade, among which 70% were published in the last five years. The remainder of this paper is organized as follows: In Section 2, various image generation tasks are introduced, as well as the differences in their evaluation. In Section 3, six common and important evaluation aspects are identified, and new subjective and objective evaluation aspect protocols are proposed for different image generation tasks. In Section 4, human evaluation methods, tools, details, and data analysis methods are presented. Classical and emerging evaluation metrics, as well as benchmarks for automatic evaluation, are reviewed in Section 5. In Section 6, challenges and future directions for image generation evaluation are explored. Section 7 concludes the paper. Fig. 2 shows the main content of this paper.

2 Various image generation tasks

Image generation tasks (Alqahtani et al., 2021; Elasri et al., 2022; Croitoru et al., 2023) are summarized based on input conditions. Content-based classification such as classification of face and medical images is not discussed because expertise is required. Fig. 3 summarizes 10 classification tasks.

2.1 Image-to-image generation

Image-to-image generation involves translating an image from the source domain to the target domain while preserving the content representation of the original image as much as possible (Pang et al., 2022) by learning a mapping function (Ak et al., 2019).

This process comprises various sub-tasks (Pang et al., 2022), among which 10 important ones have been summarized with focus on their evaluation aspects: image super-resolution (Odena et al., 2017; Wang ZH et al., 2021), image inpainting (Li JY et al., 2020; Quan et al., 2022; Zhang XB et al., 2023), style transfer (Luan et al., 2017; Ioannou and Maddock, 2024), image-to-cartoon generation (Liu YF et al., 2018; Dong et al., 2022; Zhao Y et al., 2022), image colorization (Zhang R et al., 2016; Nazeri et al., 2018), attribute manipulation (Shen and Liu, 2017; Ak et al., 2019; Chen YC et al., 2019), semantic manipulation (Hong S et al., 2018; Liang XD et al., 2018; Li G et al., 2022), image dehazing (Li BY et al., 2019; Zhao SY et al., 2021), image deblurring (Tao et al., 2018; Zhang HG et al., 2019; Cho SJ et al., 2021; Zhang KH et al., 2022; Zheng BY et al., 2024), and low-light enhancement (Fan et al., 2020; Jin et al., 2021; Liu RS et al., 2022; Liu JX et al., 2024).

Existing image-to-image generation tasks use automatic metrics, such as the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), and quality evaluation metrics, such as the Fréchet inception distance (FID) and learned perceptual image

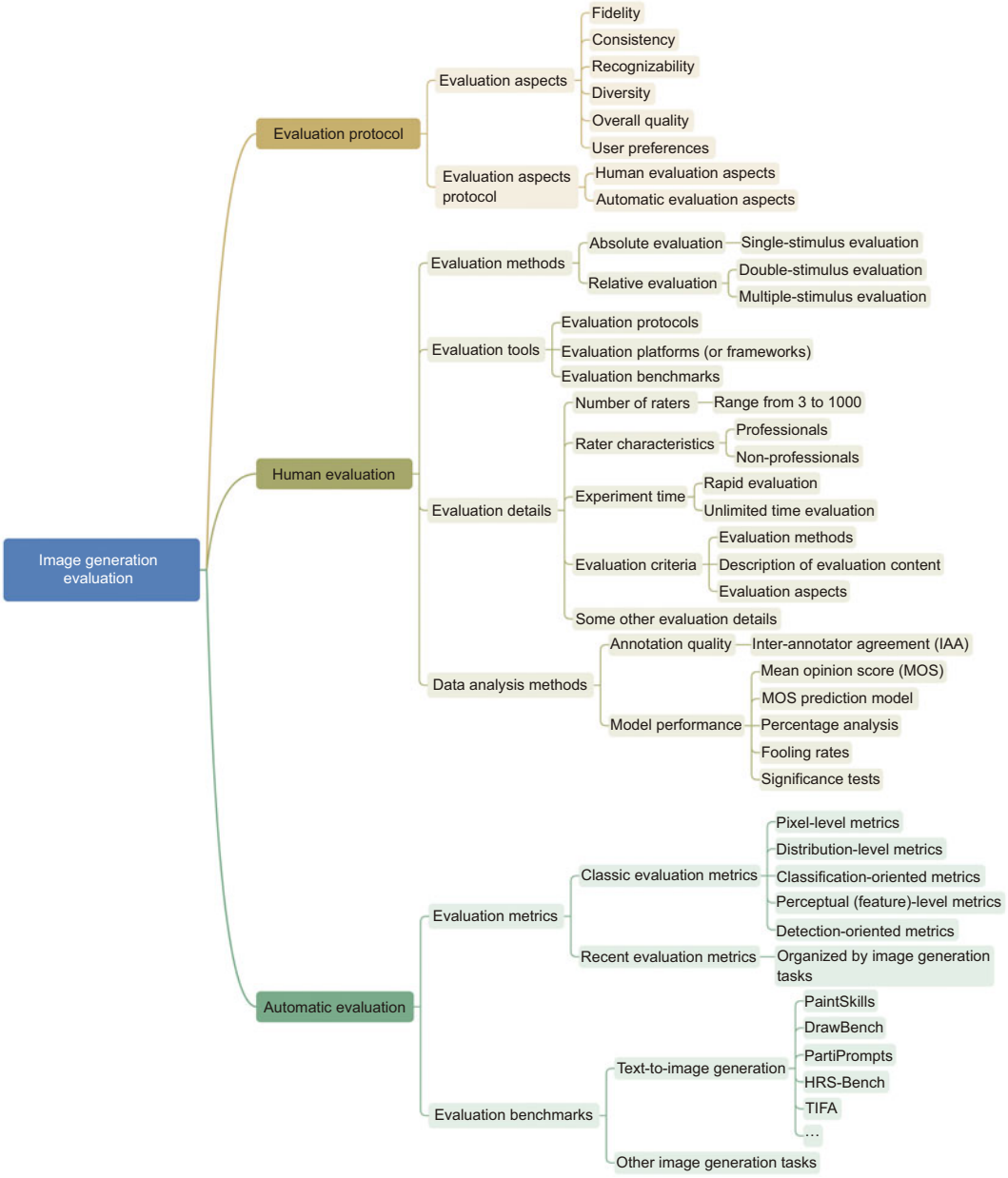

Fig. 2 Main scope of image generation evaluation in this paper

patch similarity (LPIPS), to assess the quality of generated images, but user studies are relatively rare. However, some automatic evaluation metrics show discrepancies with human perception in some image processing tasks, indicating potential limitations of relying solely on these metrics. For instance, Liu JX et al. (2024) pointed out that SSIM and PSNR are insufficient for evaluating image dehazing and low-light enhancement, as they fail to adequately reflect human visual perception. Image quality remains the primary evaluation aspect, encompassing both objective quality measured by automatic metrics and subjective quality assessed through human perception and overall perception. In addition, fidelity, consistency, recognizability, user preference, and diversity are important evaluation aspects.

2.2 Sketch-to-image generation

Sketch is an abstract visual representation, and sketch-to-image generation involves converting an input hand-drawn sketch into a realistic image. Wu

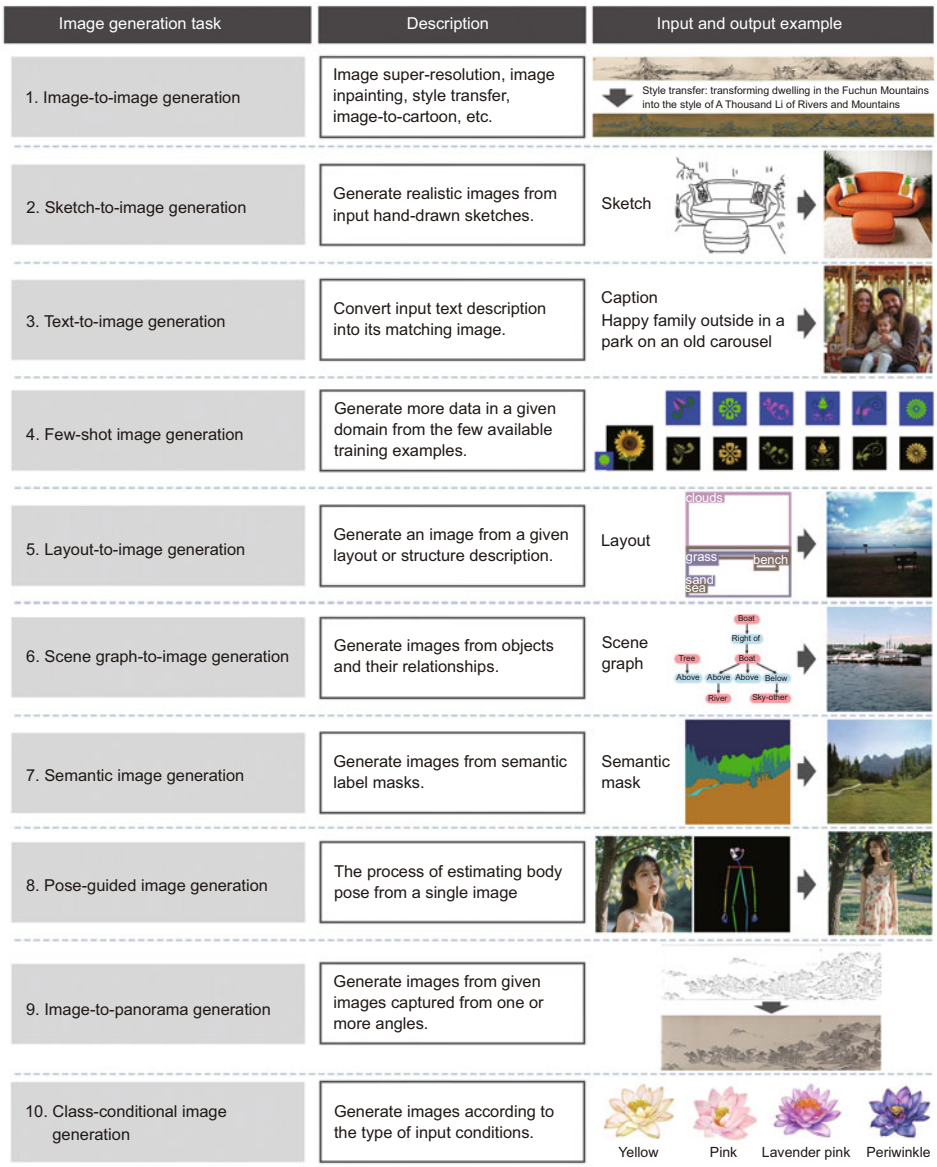

Fig. 3 Summary of image generation tasks

ZB et al. (2023) reported that sketch-to-image generation models must accurately interpret the input sketch to recognize object shapes and categories and generate realistic images. Sketch-to-image generation task involves generating a single object (e.g., a cat) and complex scenes from multiple objects.

Fidelity (Gao CY et al., 2020), consistency (Gao CY et al., 2020), diversity (Koley et al., 2023), and overall quality (Ho TT et al., 2020; Xia et al., 2021a; Koley et al., 2023; Wu ZB et al., 2023) are key aspects for automatic evaluation, whereas fidelity (Gao CY et al., 2020), consistency (Gao CY et al., 2020; Koley et al., 2023), and overall quality (Ho TT et al., 2020; Wu ZB et al., 2023) are key aspects for human evaluation.

2.3 Text-to-image generation

Text-to-image generation involves converting an input text into a matching image. Text is more flexible because it can describe a wide range of concepts and levels of details compared to other inputs (Zhang H et al., 2021), but it is also highly ambiguous, making it challenging to enforce constraints in complex generative scenarios (Zhao B et al., 2019). Compared with inputs such as layout diagrams, text may not

accurately represent the size and positional relationships of objects.

Moreover, text alignment is a major challenge in text-to-image generation, but it is an important evaluation aspect. For text-to-image generation, the model must have a deep understanding of the input text and generated images (Otani et al., 2023). The model handles data transformation of two different attributes, namely visual and linguistic, by learning from unstructured descriptions (Zhang H et al., 2021).

Fidelity (Yan et al., 2022; He YT et al., 2023; Otani et al., 2023), consistency (Qiao et al., 2019a, 2019b; Hinz et al., 2022; Yan et al., 2022; Otani et al., 2023; Zheng WD et al., 2024), diversity (Qiao et al., 2019a, 2019b; Xia et al., 2021b), recognizability (Zhang H et al., 2021; Hinz et al., 2022; Ramesh et al., 2022; Yan et al., 2022; He YT et al., 2023), overall quality (Xia et al., 2021b; Zhang H et al., 2021; Ding et al., 2022; Hinz et al., 2022; Kim et al., 2022; Ramesh et al., 2022; Zhou YF et al., 2023), and user preference (Zheng WD et al., 2024) are important aspects for automatic evaluation. In contrast, fidelity (Qiao et al., 2019a, 2019b; Xia et al., 2021b; Zhang H et al., 2021; Ramesh et al., 2022; Yan et al., 2022; Otani et al., 2023; Zhou YF et al., 2023), consistency (Qiao et al., 2019a, 2019b; Xia et al., 2021b; Zhang H et al., 2021; Ding et al., 2022; Kim et al., 2022; Ramesh et al., 2022; Yan et al., 2022; Otani et al., 2023; Zhou YF et al., 2023; Zheng WD et al., 2024), diversity (Ramesh et al., 2022), overall quality (Ding et al., 2022), recognizability (He YT et al., 2023), and user preference (Ding et al., 2022) are important aspects for human evaluation.

2.4 Few-shot image generation

Few-shot image generation is a task that generates new data from limited training examples (Li YJ et al., 2020), and only one sample is used in one-shot generation (Hong Y et al., 2020). It addresses scenarios with limited data (Li YJ et al., 2020). For instance, it can be applied to replicate an artist's unique style from a few existing works, generating variations of a handcrafted design, or expanding a small set of concept sketches into a diverse collection of artworks. In such cases, artists may not be able to manually create thousands of works. Additionally, generating diverse images while preserving the source domain characteristics is difficult. With limited images, models tend to overfit or produce low-quality results (Ojha et al., 2021).

Gu et al. (2021) introduced three main approaches to few-shot image generation, namely transformation-, optimization-, and fusion-based methods. Transformation-based methods are suitable for simple generation tasks such as intra-category image transformation. Optimization-based methods introduce a meta-learning paradigm to learn an initialization policy for an unconditional image generation task that can be quickly adapted to new few-shot tasks. Fusion-based methods combine the features of multiple input images to generate more realistic and diverse images.

In the automatic evaluation of these models, diversity (Li YJ et al., 2020; Gu et al., 2021; Ojha et al., 2021; Phaphuangwittayakul et al., 2022; Wang YH et al., 2022; Xie et al., 2022; Li LX et al., 2023; Zhao YQ et al., 2023; Zhu JY et al., 2024), overall quality (Hong Y et al., 2020; Li YJ et al., 2020; Gu et al., 2021; Ojha et al., 2021; Phaphuangwittayakul et al., 2022; Wang YH et al., 2022; Zhao YQ et al., 2023; Zhu JY et al., 2024), and fidelity (Li LX et al., 2023) are considered key evaluation aspects. In human evaluation, fidelity (Li YJ et al., 2020), overall quality (Xie et al., 2022; Zhu JY et al., 2024), and diversity (Xie et al., 2022) are considered key evaluation aspects.

2.5 Layout-to-image generation

Layout-to-image generation is a task that generates images from layouts or structural descriptions (Cheng et al., 2023). A layout is a segmentation mask or a collection of labeled object-bounding boxes (Cheng et al., 2023). It offers flexibility and control for generating complex scenes (Sylvain et al., 2021; Cheng et al., 2023) and has a user-friendly input format. It can also be used as an intermediate step for tasks such as text-to-image generation and scene graph image generation (He S et al., 2021).

However, layout-to-image generation faces the problems of one-to-many mapping (e.g., the same layout can result in many plausible implementations) and handling consistent multiobject generation and uneven occlusion relationships (e.g., bounding boxes and labels convey limited information and may encounter object interactions and occlusion relationships between objects in overlapping bounding boxes) (Zhao B et al., 2019; Sun and Wu, 2021).

Layout-to-image generation should meet three key requirements: fidelity, recognizability, and diversity (Zhao B et al., 2019). Its quality and user preference have also been evaluated. In automatic evaluation, fidelity (Zhao B et al., 2019; Sylvain et al., 2021), consistency (Li ZJ et al., 2021; Sylvain et al., 2021; Cheng et al., 2023), recognizability (Zhao B et al., 2019; Li ZJ et al., 2021; Sylvain et al., 2021; Wu JY et al., 2022; Cheng et al., 2023), overall quality (Zhao B et al., 2019; He S et al., 2021; Li ZJ et al., 2021; Sun and Wu, 2021; Sylvain et al., 2021; Wu JY et al., 2022; Cheng et al., 2023), and diversity (Zhao B et al., 2019; He S et al., 2021; Li ZJ et al., 2021; Sun and Wu, 2021; Sylvain et al., 2021; Wu JY et al., 2022; Cheng et al., 2023) are key evaluation aspects. In human evaluation, user preference (He S et al., 2021; Li ZJ et al., 2021), consistency (Sylvain et al., 2021; Wu JY et al., 2022), fidelity (Wu JY et al., 2022; Cheng et al., 2023), and overall quality (Cheng et al., 2023) are key evaluation aspects.

2.6 Scene graph-to-image generation

Scene graph-to-image generation is a task that generates images from objects and their relationships using various intermediate representations such as layouts (Hassan et al., 2023), masks (Farshad et al., 2023; Wang ZM et al., 2023), bounding boxes (Farshad et al., 2023; Zhang YK et al., 2023), and embeddings of points and edges (Liu JX et al., 2024). Although this task uses different frameworks, its common goal is to decompose the overall task into subtasks: abstract semantic understanding, relational consistency, entity location estimation, entity generation, and combined optimization for image generation; these tasks are performed to produce high-quality images. However, effective selection and optimization of intermediate representations remain challenging. Given the complexity of scene graph-to-image generation and the need for multistep composite optimization, the model must integrate scene graph understanding, maintain entity relationships, and accurately generate detailed images.

Contrary to text, scene graphs offer a powerful structured representation that graphically depicts objects, attributes, and their relationships (Tripathi et al., 2019b; Zhao B et al., 2019; Hua et al., 2021). Scene graphs are a robust tool for visual scene understanding (Chang et al., 2023; Hassan et al., 2023) and are widely applied in tasks such as image retrieval and image captioning.

In the automatic evaluation of these models, fidelity (Hua et al., 2021; Wang ZM et al., 2023; Zhang YK et al., 2023), recognizability (Johnson et al., 2018; Hassan et al., 2023; Zhang YK et al., 2023), diversity (Johnson et al., 2018; Luo et al., 2020; Hassan et al., 2023; Zhang YK et al., 2023), overall quality (Johnson et al., 2018; Hua et al., 2021; Hassan et al., 2023; Wang ZM et al., 2023; Zhang YK et al., 2023), and consistency (Tripathi et al., 2019a, 2019b; Luo et al., 2020) are considered key evaluation aspects. In human evaluation, the key evaluation aspects are consistency (Johnson et al., 2018; Tripathi et al., 2019b), recognizability (Johnson et al., 2018; Zhang YK et al., 2023), diversity (Luo et al., 2020), and user preference (Hua et al., 2021; Zhang YK et al., 2023).

2.7 Semantic image generation

Semantic image generation is a simplified task that generates an image from a semantic label mask (Tan et al., 2021) by rendering realistic images according to user-specified layouts without requiring complex graphics engines (Sushko et al., 2021). Semantic image generation provides a simple and controllable way to improve the consistency of generation results for many other image generation tasks.

In semantic multimodal image generation, diverse natural images are generated based on the semantic labels of the input. This approach focuses on realizing multimodality at the semantic level; i.e., the controller of a specific semantic area is adjusted when generating images, and its diversified images are generated while keeping other semantic areas unchanged (Zhu Z et al., 2020).

However, preventing the removal of semantic information during propagation remains a key challenge in semantic image generation (Chen P et al., 2022). Semantic consistency is a key evaluation aspect for semantic image generation. Specifically, consistency (Park et al., 2019; Habtegebrial et al., 2020; Tang H et al., 2020a; Zhu Z et al., 2020; Sushko et al., 2021; Tan et al., 2021; Chen P et al., 2022; Tang H et al., 2023), overall quality (Park et al., 2019; Tang H et al., 2020a; Sushko et al., 2021; Tan et al., 2021; Chen P et al., 2022; Tang H et al., 2023), diversity (Zhu Z et al., 2020; Sushko et al., 2021; Tan et al., 2021), and fidelity (Habtegebrial et al., 2020; Zhu Z et al., 2020) are key aspects for

automatic evaluation of such models, whereas user preference (Park et al., 2019; Tang H et al., 2020a; Chen P et al., 2022; Tang H et al., 2023), consistency (Zhu Z et al., 2020), fidelity (Tan et al., 2021; Tang H et al., 2023), and diversity (Tan et al., 2021) are key aspects for human evaluation.

2.8 Pose-guided image generation

Pose-guided image generation is a task that generates images based on the estimated body pose from a single (usually monocular) image. It aims to transform images of people from source poses to target poses while preserving their appearance details (Shi et al., 2022; Chen JX et al., 2023). However, the generated images must be as photo-realistic as possible (Lv et al., 2021). Pose-guided image generation has broad application prospects in areas such as virtual try-on, film production, e-commerce, and virtual reality (Zhang PZ et al., 2022; Chen JX et al., 2023).

Pose-guided image generation involves spatial processing of source data. The target image can be viewed as a non-rigidly deformed version of the source image, and the image of the human body must undergo complex geometric deformation before implementation (Ren et al., 2020; Lv et al., 2021). Generating images of people in target poses remains a challenging task due to issues such as complex texture reconstruction, spatial arrangement, geometric deformation, inference of self-occlusion and invisible areas, and realism and consistency of generation results (Lv et al., 2021; Shi et al., 2022). Human pose transfer also faces challenges in handling deformable humans, complete and partial views, and clothing details (Lu et al., 2022).

Pose-guided image generation also requires that the generated images be similar to the target images. In the automatic evaluation of these models, overall quality (Ma et al., 2017; Siarohin et al., 2018; Zhu Z et al., 2019; Ren et al., 2020; Tang H et al., 2020b; Lv et al., 2021; Lu et al., 2022; Shi et al., 2022; Wang ZJ et al., 2022; Zhang PZ et al., 2022), consistency (Zhu Z et al., 2019; Tang H et al., 2020b; Lv et al., 2021; Zhang PZ et al., 2022), diversity (Lv et al., 2021), and fidelity (Ren et al., 2020; Lv et al., 2021) are key evaluation aspects. In human evaluation, fidelity (Ma et al., 2017; Siarohin et al., 2018; Zhu Z et al., 2019; Ren et al., 2020; Tang H et al., 2020b; Lv et al., 2021; Lu et al., 2022; Zhang PZ et al., 2022) and overall quality (Lv et al., 2021) are considered key evaluation aspects.

2.9 Image-to-panorama generation

Panorama generation is a task that generates panoramic images from a text prompt, a single image, or given images captured from one or more angles. Panoramic images such as scroll paintings or 360° images comprise spherical images captured from different viewpoints (Duan et al., 2020). Panoramic images and traditional image generation models are considerably different because panoramic images provide users with an immersive and interactive viewing experience; they allow users to freely switch viewing angles within the range of $360° \times 180°$ (Duan et al., 2020). These images can be widely used across fields such as surveillance systems, construction, tourism, self-driving cars, and entertainment (Hara et al., 2021).

Multidiffusion (Bar-Tal et al., 2023) uses a pretrained text-to-image diffusion model to generate different regions of a panorama; these are then averaged to create a panorama from a text prompt. CamFreeDiff (Yuan et al., 2024) takes a single camera-free image and text description as the input and estimates the camera position from the image to produce a 360° panorama. PanFusion (Zhang C et al., 2024), a dual-branch diffusion model, generates 360° images from text prompts using a unique cross-attention mechanism. However, long scroll generation poses challenges such as repetition, omission, or illogicality of object layouts, as well as incoherence of scene layouts between different perspectives (Shibata et al., 2014; Cai et al., 2024).

Fidelity (Duan et al., 2020), consistency (Zhang JM et al., 2022), and overall quality (Hara et al., 2021; Cai et al., 2024) are key aspects for automatic evaluation. Overall quality (Shibata et al., 2014), and user preference (Duan et al., 2020; Marrinan and Papka, 2021; Cai et al., 2024) are key aspects for human evaluation.

2.10 Class-conditional image generation

Class-conditional image generation is a basic conditional image generation task, wherein images containing a specified class are generated by conditioning a generative model on a class label (such as “dog” or “cat”) (Odena et al., 2017; Foo et al.,

2023). This task ensures that the generated images have realistic visuals and clear class-specific features, thereby improving their quality and diversity.

In automatic evaluation, fidelity (Sauer et al., 2022), recognizability (Odena et al., 2017; Saseendran et al., 2021), diversity (Odena et al., 2017; Brock et al., 2019; Sauer et al., 2022), and overall quality (Odena et al., 2017; Brock et al., 2019; Huh et al., 2020; Saseendran et al., 2021; Ho J et al., 2022; Sauer et al., 2022) are used as the evaluation aspects. Contrarily, fidelity (Huh et al., 2020) is considered an important human evaluation aspect.

3 Protocol for evaluation aspects

The evaluation metrics vary across different image generation tasks (Section 2). However, the aspects to be measured, which we refer to as evaluation aspects, remain largely the same. Specifically, we summarize the following two scenarios:

1. In automatic evaluation, different metrics are used to assess the same aspect. For instance, in few-shot image generation, Zhao YQ et al. (2023) used LPIPS to evaluate image diversity, whereas Xie et al. (2022) used FID; however, they both assessed diversity. Moreover, similar scenarios occur in different image generation tasks. For semantic consistency, Chen P et al. (2022) used mean intersection-over-union (mIoU) and pixel accuracy to measure consistency between the input semantic map and generated image. Contrarily, Qiao et al. (2019a) used R-Precision to evaluate the consistency between the input text and the generated image.

2. In human evaluation, descriptions of evaluated aspects may vary but the underlying criteria are similar. For instance, in sketch-to-image generation, Gao CY et al. (2020) asked participants to assess faithfulness by selecting the image that was most likely generated from a given sketch, whereas Koley et al. (2023) asked participants to rate the level of matching between sketches and generated images. In text-to-image generation, Xia et al. (2021b) evaluated accuracy by asking participants to choose images that were most coherent with the input text. These studies collectively assessed the consistency between input content (sketch or text) and the generated image.

Inspired by these studies, we extracted six common important evaluation aspects (Section 3.1) to have a more systematic understanding of the evaluation metrics for each image generation task. We then summarized automatic and human evaluation metrics used for each image generation task and proposed a protocol including the subjective and objective evaluation aspects (Section 3.2).

3.1 Introduction to evaluation aspects

We extracted six evaluation aspects: fidelity, consistency, recognizability, diversity, overall quality, and user preference. Their reference sources are shown in Table 1, which revealed that fewer studies used human evaluation and that evaluation methods are relatively limited.

Human evaluation in existing research has three main types of annotations:

1. Providing a detailed description of the aspect to be assessed via textual descriptions (Qiao et al., 2019b; Xia et al., 2021b; Zhang H et al., 2021; Kim et al., 2022; Ramesh et al., 2022). For instance, Xia et al. (2021b) evaluated accuracy and realism via user studies by asking users to judge images that are more realistic and consistent with a given text.

2. Asking participants to rate several other relevant evaluation aspects for the aspect being assessed (Joo et al., 2018; He S et al., 2021; Wu YZ et al., 2021; Cai et al., 2024). For example, He S et al. (2021) asked users to consider two aspects for obtaining their preference: image quality and layout matching; they then selected their favorite image.

3. Providing no explanation and asking participants to rate generated images directly (Tang H et al., 2020a; Xie et al., 2022; Yan et al., 2022; Zhou YF et al., 2023; Zheng WD et al., 2024). Zhou YF et al. (2023) required users to make judgments about alignment and fidelity without providing any other instructions. However, the wording of the questions can affect the outcomes of evaluation (Ioannou and Maddock, 2024). The differences in annotations for these evaluation aspects may lead to inconsistent experimental designs and unfair comparisons, ultimately resulting in inaccurate evaluation results. To this end, we have provided clear and standardized annotations on the evaluation aspects in Table 2.

3.2 Proposed protocol for evaluation aspects

Evaluation aspects vary across different image generation tasks; therefore, appropriate evaluation

Table 1 Evaluation aspect refinement

Evaluation aspect	Automatic evaluation metric	Human evaluation content
Fidelity	FID (Heusel et al., 2017): Zhao B et al., 2019; Duan et al., 2020; Gao CY et al., 2020; Habtegebrial et al., 2020; Ren et al., 2020; Zhu Z et al., 2020; Lv et al., 2021; Yan et al., 2022; Li LX et al., 2023; Otani et al., 2023 Accuracy (Ashual and Wolf, 2019): Gao CY et al., 2020 SSIM (Wang Z et al., 2004): Gao CY et al., 2020 Inception score (IS) (Salimans et al., 2016): Duan et al., 2020; Hua et al., 2021; Yan et al., 2022; Wang ZM et al., 2023 Kernel inception distance (KID) (Bińkowski et al., 2021): He YT et al., 2023 Perceptual distance (Zhang R et al., 2018): Habtegebrial et al., 2020 Precision (Kynkäänniemi et al., 2019): Sauer et al., 2022 SceneFID (Sylvain et al., 2021): Zhang YK et al., 2023	Realistic: Qiao et al., 2019b; Gao CY et al., 2020; Li YJ et al., 2020; Xia et al., 2021b; Zhang H et al., 2021; Zhang PZ et al., 2022 Photorealism: Ramesh et al., 2022 Fidelity: Wu JY et al., 2022; Yan et al., 2022; Cheng et al., 2023; Otani et al., 2023; Zhou YF et al., 2023 Real & fake: Ma et al., 2017; Siarohin et al., 2018; Qiao et al., 2019a; Zhu Z et al., 2019; Huh et al., 2020; Ren et al., 2020; Tang H et al., 2020b; Lv et al., 2021; Lu et al., 2022
Consistency	Shape similarity (SS) (Gao CY et al., 2020) R-Precision (Xu T et al., 2018): Qiao et al. 2019a, 2019b; Hinz et al., 2022 Spatial semantic CLIP score (Yan et al., 2022) CLIP score (Hessel et al., 2022): Otani et al., 2023 Human preference score v2 (HPS v2) (Wu XS et al., 2023b): Zheng WD et al., 2024 ImageReward (Xu JZ et al., 2023): Zheng WD et al., 2024 YOLO score (Li ZJ et al., 2021): Cheng et al., 2023 SceneFID (Sylvain et al., 2021): Li ZJ et al., 2021; Cheng et al., 2023 mIoU: Park et al., 2019; Habtegebrial et al., 2020; Tang H et al., 2020a; Zhu Z et al., 2020; Sushko et al., 2021; Tan et al., 2021; Chen P et al., 2022; Zhang JM et al., 2022; Cheng et al., 2023; Tang H et al., 2023 Relation score (Tripathi et al., 2019b): Tripathi et al., 2019a Intersection-over-union (IoU): Tripathi et al., 2019a Accuracy: Luo et al., 2020 Pixel accuracy: Park et al., 2019; Tang H et al., 2020a; Zhu Z et al., 2020; Tan et al., 2021; Chen P et al., 2022; Tang H et al., 2023 Class accuracy: Habtegebrial et al., 2020 L1 loss: Luo et al., 2020 PCKh (Andriluka et al., 2014): Zhu Z et al., 2019; Tang H et al., 2020b; Lv et al., 2021; Zhang PZ et al., 2022 Mean pose distances (MPD) and mean missed detections: Zhang PZ et al., 2022 Rank-k: Zhang PZ et al., 2022 Mean average precision: Zhang PZ et al., 2022	Faithfulness: Gao CY et al., 2020 Matching degree: Johnson et al., 2018; Koley et al., 2023 Alignment: Zhang H et al., 2021; Wu JY et al., 2022; Otani et al., 2023; Zhou YF et al., 2023; Zheng WD et al., 2024 Accuracy: Xia et al., 2021b; Kim et al., 2022 Semantic consistency: Qiao et al., 2019b Caption similarity: Ramesh et al., 2022 Semantic: Yan et al., 2022 Relevance: Ding et al., 2022 Pairwise test: Qiao et al., 2019a Layout-fidelity: Sylvain et al., 2021 Mean opinion relation score: Tripathi et al., 2019b SMIS human evaluation (SHE): Zhu Z et al., 2020
Recognizability	Semantic object accuracy (Hinz et al., 2022): Zhang H et al., 2021; Yan et al., 2022 Recall: Ramesh et al., 2022; He YT et al., 2023 Precision: He YT et al., 2023 mAP: Wu JY et al., 2022; He YT et al., 2023 mAP50: Wu JY et al., 2022; He YT et al., 2023 mAP75: Wu JY et al., 2022 IoU: Wu JY et al., 2022 Classification accuracy: Odena et al., 2017; Zhao B et al., 2019; Li ZJ et al., 2021; Sylvain et al., 2021; Cheng et al., 2023; Hassan et al., 2023 Object recall: Johnson et al., 2018 Average count accuracy (ACA): Saseendran et al., 2021 Object occurrence ratio (Zhang YK et al., 2023)	Accuracy: He YT et al., 2023 Object recall: Johnson et al., 2018; Zhang YK et al., 2023

To be continued

Table 1 (continued)

Evaluation aspect	Automatic evaluation metric	Human evaluation content
Diversity	LPIPS (Zhang R et al., 2018): Zhao B et al., 2019; Li YJ et al., 2020; Zhu Z et al., 2020; Gu et al., 2021; He S et al., 2021; Li ZJ et al., 2021; Lv et al., 2021; Ojha et al., 2021; Sun and Wu, 2021; Sushko et al., 2021; Tan et al., 2021; Xia et al., 2021b; Phaphuangwittayakul et al., 2022; Wu JY et al., 2022; Koley et al., 2023; Li LX et al., 2023; Zhang YK et al., 2023; Zhao YQ et al., 2023; Zhu JY et al., 2024 IS (Salimans et al., 2016): Brock et al., 2019; Qiao et al., 2019a, 2019b; Zhao B et al., 2019; Duan et al., 2020; Sylvain et al., 2021; Phaphuangwittayakul et al., 2022; Cheng et al., 2023; Hassan et al., 2023; Cai et al., 2024 FID (Heusel et al., 2017): Brock et al., 2019; Sushko et al., 2021; Sylvain et al., 2021; Sauer et al., 2022; Wang YH et al., 2022; Xie et al., 2022 Variety: Johnson et al., 2018 Diversity score: Hassan et al., 2023 Standard deviation: Luo et al., 2020 Improved precision and recall (Kynkäänniemi et al., 2019): Sushko et al., 2021 MS-SSIM (Wang Z et al., 2003): Odena et al., 2017; Sushko et al., 2021 mCSD, mOCD (Zhu Z et al., 2020): Tan et al., 2021 mISD, mOID (Tan et al., 2021) Recall (Kynkäänniemi et al., 2019): Sauer et al., 2022	Diversity: Luo et al., 2020; Ramesh et al., 2022; Xie et al., 2022
Overall quality	FID (Heusel et al., 2017): Brock et al., 2019; Park et al., 2019; Ho TT et al., 2020; Hong Y et al., 2020; Li YJ et al., 2020; Tang H et al., 2020a; Gu et al., 2021; Hara et al., 2021; He S et al., 2021; Hua et al., 2021; Li ZJ et al., 2021; Ojha et al., 2021; Saseendran et al., 2021; Sun and Wu, 2021; Sushko et al., 2021; Sylvain et al., 2021; Tan et al., 2021; Xia et al., 2021a, 2021b; Zhang H et al., 2021; Chen P et al., 2022; Ding et al., 2022; Hinz et al., 2022; Ho J et al., 2022; Kim et al., 2022; Lu et al., 2022; Ramesh et al., 2022; Phaphuangwittayakul et al., 2022; Wang YH et al., 2022; Wang ZJ et al., 2022; Wu JY et al., 2022; Xie et al., 2022; Zhang PZ et al., 2022; Cheng et al., 2023; Hassan et al., 2023; Koley et al., 2023; Tang H et al., 2023; Wang ZM et al., 2023; Wu ZB et al., 2023; Zhang YK et al., 2023; Zhao YQ et al., 2023; Zhou YF et al., 2023; Zhu JY et al., 2024 LPIPS (Zhang R et al., 2018): Hong Y et al., 2020; Huh et al., 2020; Ren et al., 2020; Lu et al., 2022; Wang ZJ et al., 2022; Zhang PZ et al., 2022; Wu ZB et al., 2023 Mask-LPIPS (MLPIPS) (Ma et al., 2017): Lu et al., 2022 IS (Salimans et al., 2016): Ma et al., 2017; Odena et al., 2017; Johnson et al., 2018; Siarohin et al., 2018; Brock et al., 2019; Zhao B et al., 2019; Zhu Z et al., 2019; Hong Y et al., 2020; Tang H et al., 2020b; He S et al., 2021; Sun and Wu, 2021; Sylvain et al., 2021; Zhang H et al., 2021; Ding et al., 2022; Hinz et al., 2022; Ho J et al., 2022; Sauer et al., 2022; Shi et al., 2022; Wang ZJ et al., 2022; Wu JY et al., 2022; Cheng et al., 2023; Hassan et al., 2023; Wu ZB et al., 2023; Zhang YK et al., 2023; Zhou YF et al., 2023; Cai et al., 2024 Fine-grained metric (FGM) (Koley et al., 2023) Peak signal-to-noise ratio (PSNR): Hara et al., 2021; Xia et al., 2021a; Zhang PZ et al., 2022	Perceived quality: Ho TT et al., 2020; Wu ZB et al., 2023 Overall quality: Lv et al., 2021; Ding et al., 2022; Lu et al., 2022; Cheng et al., 2023; Zhu JY et al., 2024 Quality: Shibata et al., 2014; Xie et al., 2022

To be continued

Table 1 (continued)

Evaluation aspect	Automatic evaluation metric	Human evaluation content
Overall quality	SSIM (Wang Z et al., 2004): Ma et al., 2017; Siarohin et al., 2018; Zhu Z et al., 2019; Ho TT et al., 2020; Tang H et al., 2020b; Lv et al., 2021; Xia et al., 2021a; Shi et al., 2022; Wang ZJ et al., 2022; Zhang PZ et al., 2022 mask-SSIM (Ma et al., 2017): Siarohin et al., 2018; Zhu Z et al., 2019; Tang H et al., 2020b; Shi et al., 2022 mask-IS (Ma et al., 2017): Siarohin et al., 2018; Zhu Z et al., 2019; Tang H et al., 2020b; Shi et al., 2022 R-Precision: Zhang H et al., 2021 KID (Li HL et al., 2018): Zhao YQ et al., 2023 Precision and recall (Sajjadi et al., 2018): Zhu JY et al., 2024 Classification accuracy score (CAS) (Ravuri and Vinyals, 2019): Sun and Wu, 2021 Improved precision and recall (Kynkäänniemi et al., 2019): Sushko et al., 2021 Detection score (Siarohin et al., 2018): Zhu Z et al., 2019; Shi et al., 2022 Part-based SSIM (PSSIM) (Shi et al., 2022) L1 error: Hara et al., 2021 Per-pixel: Huh et al., 2020	
User preference	HPS v2 (Wu XS et al., 2023b): Zheng WD et al., 2024 ImageReward (Xu JZ et al., 2023): Zheng WD et al., 2024	User preference: Park et al., 2019; Duan et al., 2020; Tang H et al., 2020a; He S et al., 2021; Hua et al., 2021; Marrinan and Papka, 2021; Ding et al., 2022; Zhang YK et al., 2023; Cai et al., 2024

Table 2 Evaluation aspects and annotations

Evaluation aspect	Annotation
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 preference and inclination towards images

aspects must be chosen. We have summarized 10 image generation tasks and analyzed the automatic metrics and human evaluation content of each task in existing studies. We have also provided a subjective and objective evaluation aspect protocol for various image generation tasks (Table 3). Results revealed that some tasks have only automatic evaluation for certain evaluation aspects (e.g., layout-to-image generation measures recognizability and diversity using only automatic evaluation). In contrast, user preferences were obtained mainly via human evaluation.

This protocol is based on the summary of previous studies. We suggest that researchers refer to this protocol and adjust it based on their specific research objectives when determining the evaluation aspects. We hope to continue to update these evaluation aspects as generative models undergo further advancement.

4 Human evaluation

Human evaluation involves assessing the image quality via human subjective perception and is considered the most direct and reliable method (Xu QQ et al., 2012; Ma and Fang, 2021; Wang J et al., 2023). Automatic evaluation is useful but does not fully reflect human perception (Xu JZ et al., 2023; Cai et al., 2024; Liu JX et al., 2024). The evaluation of image generation models is challenging and distinct from traditional tasks such as image classification or detection. This is because it involves aspects that are difficult to quantify such as quality, aesthetics, and

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

faithfulness to the text prompt (Jayasumana et al., 2024). The evaluation of generated images inherently involves human perception and understanding, making human evaluation essential (Xu QQ et al., 2012; Otani et al., 2023).

Many studies (Frolov et al., 2021; Otani et al., 2023; Ioannou and Maddock, 2024) have introduced human evaluation and highlighted the need for using a standardized protocol. The implementation of standardized human evaluation has numerous advantages, including but not limited to the following points: (1) Standardized evaluation allows for the direct comparison of results across different studies, offering good generalizability; (2) Standardized evaluation typically contributes to the reliability and reproducibility of experimental results; (3) Standardization helps reduce subjective biases in user experiments, thereby ensuring the objectivity of results. However, varying protocols across studies have not been systematically compared or summarized, thereby hindering the consensus and development of standard practices. We attempted to conduct a comprehensive summary analysis of human evaluation for the first time, including evaluation methods, evaluation tools, evaluation details, and data analysis methods.

4.1 Evaluation methods

Human evaluation comprises two main rating methods: absolute evaluation and comparative evaluation (Xu QQ et al., 2012; Li BY et al., 2019; Khashabi et al., 2022; Otani et al., 2023). Absolute evaluation requires participants to independently assess each stimulus (e.g., an image) without comparison with other stimuli based on personal standards, perceptions, or preset scales. Comparative evaluation requires participants to compare two or more stimuli and focus on the differences, preferences, or rankings between stimuli. Fig. 4 summarizes the human evaluation methods, and Fig. 5 presents partial user interfaces corresponding to these methods.

Human evaluation can be categorized into three types based on the number of stimuli (Ma and Fang, 2021): (1) single-stimulus evaluation—participants are shown only one stimulus (e.g., an image) and are asked to rate the image based on their perception; (2) double-stimulus evaluation—two stimuli (e.g., two images, usually a reference image and a distorted image) are presented to the participants, who are required to compare these stimuli and provide a rating; (3) multiple-stimulus evaluation—multiple stimuli (such as multiple images) are presented to the participants, who are required to sort these stimuli or evaluate them based on certain criteria.

Three main evaluation methods were summarized in user studies: scale, comparative, and ranking evaluation. Scale and comparative evaluation methods are more commonly used than the ranking evaluation method.

Scale evaluation is a typical absolute evaluation

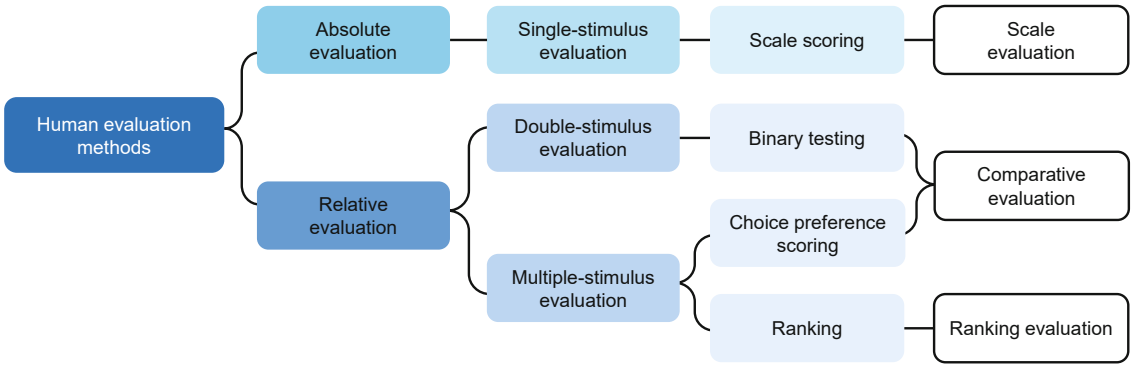

Fig. 4 Summary of human evaluation methods

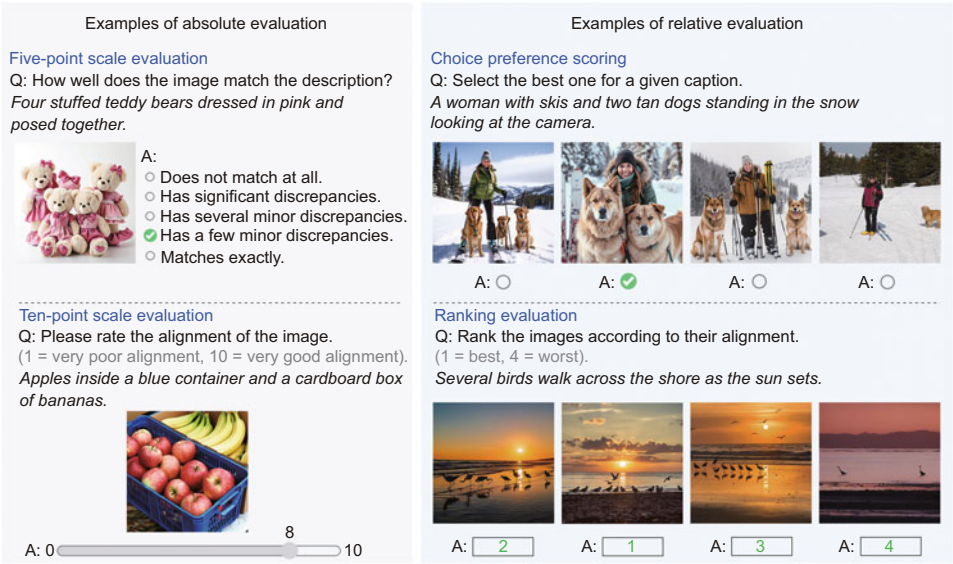

Fig. 5 User interfaces for human evaluation. This figure shows scale evaluation, comparative evaluation, and ranking evaluation under two evaluation methods: absolute evaluation and relative evaluation

method that aims to quantitatively evaluate the image quality. It usually relies on a predefined scale that contains a series of quality dimensions or criteria, against which the participants score content. Scale evaluation requires participants to score images, typically within a certain range. For example, Koley et al. (2023) asked users to rate images on a scale of 1–5, with 1 representing the worst and 5 the best.

Ranking evaluation does not involve direct scoring. Instead, this method compares the relative quality of samples. Participants were asked to rank a set of samples from the highest to lowest quality based on their subjective judgment. For example, Zhang H et al. (2021) asked participants to rank images from best to worst based on realism and alignment of samples.

Comparative evaluation (choice preference scoring or binary testing) requires participants to select one sample in pairs or sets of samples. For example, Ma et al. (2017), Siarohin et al. (2018), Zhu Z et al. (2019), Ren et al. (2020), Tang H et al. (2020b), and Marrinan and Papka (2021) required participants to judge whether an image is real or fake. Similarly, Li YJ et al. (2020), Chen P et al. (2022), Zhang PZ et al. (2022), Tang H et al. (2023), and Zhu JY et al. (2024) asked participants to select the best image from multiple images.

4.2 Evaluation tools

Human evaluation tools are divided into three categories: evaluation protocols, evaluation

platforms (or frameworks), and evaluation benchmarks.

An evaluation protocol is a guide that outlines the steps and rules for conducting an evaluation, including task instructions, evaluator training, task presentation methods, data collection methods, and result analysis steps. Otani et al. (2023) proposed a human evaluation protocol for text-to-image generation and developed a template that includes details about datasets for setup (captions, ratings/items, unique annotators, tested models, types of rating, and evaluation criteria) and annotators (platform, annotator qualifications, compensation, interface, instructions, and inter annotator agreement). Liang YW et al. (2024) provided detailed instructions covering multiple aspects such as annotation steps, interactions with the web user interface, examples of different types of implausibility, artifacts, and misalignment. Hu et al. (2023) and Lee et al. (2023) provided specific definitions for each human evaluation question and rating choice. Wu XS et al. (2023b), Xu JZ et al. (2023), and Ku et al. (2024) provided detailed instructions and listed numerous rating examples. In conclusion, the public availability and use of standardized protocols contribute to more convincing results and experimental transparency (Ku et al., 2024).

An evaluation platform (framework) is a tool or a framework for implementing the evaluation process that includes carrying out assessment missions, collecting assessment data, and conducting preliminary analyses. This platform provides an efficient operating environment that enables researchers to easily perform evaluation tasks. Amazon Mechanical Turk is a crowdsourcing platform that is widely used by researchers (Ma et al., 2017; Agustsson et al., 2019; Park et al., 2019; Li YJ et al., 2020; Shocher et al., 2020; Tang H et al., 2020a; He S et al., 2021; Otani et al., 2023) to post tasks, recruit evaluators, and collect feedback. Moreover, many researchers have developed their own evaluation frameworks. Xu QQ et al. (2012) proposed a subjective image quality evaluation framework known as HodgeRank on random graphs, which can be used for large-scale online crowdsourcing. This framework can process online crowdsourced data, obtain real-time image scores, and monitor topological changes and inconsistencies in scores in real time. Although most researchers have developed their own scoring platforms, the differences between them are not conducive to comparison between studies; these differences also hinder us from reaching generalized conclusions. Although scale evaluations have been used in all reviewed studies, Xu JZ et al. (2023) used a seven-point score range, whereas the others (Huang KY et al., 2023; Lee et al., 2023; Otani et al., 2023; Liang YW et al., 2024) used a five-point score range. Different scoring ranges can affect the consistency of scoring (Ku et al., 2024).

An evaluation benchmark is a set of standardized testing methods and datasets used to compare the performance of different models. For instance, Zhou S et al. (2019) proposed HYPE, a standardized human evaluation benchmark, which used two methods, HYPEtime and HYPE∞. This benchmark was tested and compared with four datasets: CelebA, FFHQ, CIFAR-10, and ImageNet.

Some studies have combined the aforementioned three tools to explore more comprehensive evaluation tools. Ku et al. (2024) provided a comprehensive evaluation tool, ImagenHub, which contained evaluation benchmarks, evaluation platforms, and standard protocols, ensuring a fair comparison of different models under the same conditions.

However, standard and reproducible evaluation tools are still lacking. Current evaluations conducted on human-evaluated systems are ad hoc due to a lack of unified standards and processes. Moreover, the evaluation results considerably vary depending on the details of the task design (Zhou S et al., 2019). As a unified evaluation standard does not exist, most researchers follow different protocols (Otani et al., 2023). To alleviate this confusing situation, some studies (Zhou S et al., 2019; Otani et al., 2023; Ku et al., 2024) have explored standardized human evaluation tools.

4.3 Evaluation details

The experimental results are impacted by the evaluation details of a user study. Although image quality is constant, users' focus may be influenced by factors such as the type of image generation, evaluation criteria, and experimental details during subjective evaluation, leading to different results (Xu QQ et al., 2012). We have summarized four types of evaluation details: the number of participants, participant characteristics, experiment time, and evaluation criteria. Fig. 6 provides the

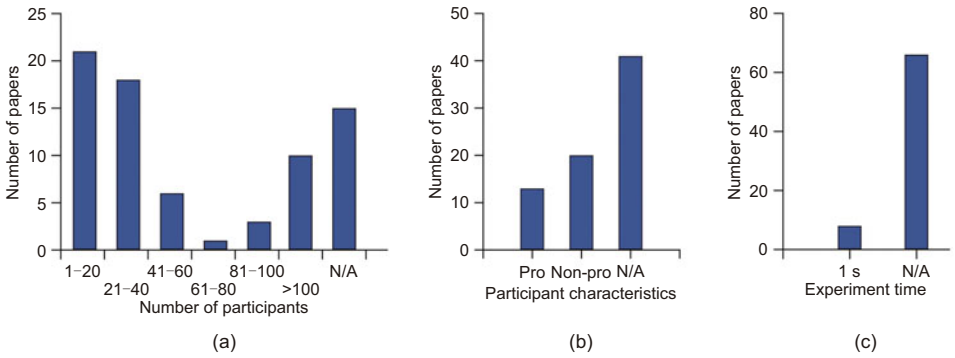

Fig. 6 We reviewed the cited references and identified 74 papers with user studies. We recorded the number of participants (a), their characteristics (b), and the experiment time (c). "N/A" denotes unspecified rater numbers, "Pro" indicates professionals, "Non-pro" denotes non-professionals, and "1 s" denotes rapid assessment within one second

number of participants, participant characteristics, and experiment time reported in existing studies.

Number of participants: the number of participants can affect the reliability of results. The number of participants recruited in the surveyed papers ranged from 3 (Johnson et al., 2018) to 1000 (Yu JH et al., 2022); however, some studies (Park et al., 2019; Huh et al., 2020; Ramesh et al., 2022; Yan et al., 2022; He YT et al., 2023; Zheng WD et al., 2024) did not specify the number of participants. Zhou S et al. (2019) suggested that recruiting more participants leads to more consistent results but increases time and cost.

Participant characteristics: individual differences among users can affect ratings, making it essential to distinguish between professionals and non-professionals (Xu QQ et al., 2012; Ioannou and Maddock, 2024). Studies on participant characteristics are categorized into three types: professionals (Siarohin et al., 2018; Guo et al., 2021; Hua et al., 2021; Lv et al., 2021; Yu YC et al., 2021; Wang J et al., 2023), non-professionals (Shibata et al., 2014; Qiao et al., 2019b; Xia et al., 2021b), and unspecified users (Johnson et al., 2018; Chen X et al., 2019; Grigorev et al., 2019; Zhu Z et al., 2019; Gao CY et al., 2020; Hulzebosch et al., 2020; Shocher et al., 2020; Volokitin et al., 2020; Yang H et al., 2020; Zhang H et al., 2021; Xie et al., 2022; Yang S et al., 2022; Koley et al., 2023). Zhu WH et al. (2018) revealed that a small number of professionals are more suitable for evaluating the perceptual quality of images than a large number of non-professionals. Hulzebosch et al. (2020) found that non-professionals can achieve higher accuracy in image recognition than random guess. However, the differential impact of professionals and non-professionals remains unclear.

Experiment time: human evaluations face an inherent trade-off between accuracy and time (Zhou S et al., 2019). Studies typically adopted two experiment time settings: rapid assessment within limited time (e.g., one second) (Ma et al., 2017; Zhu Z et al., 2019; Huh et al., 2020; Tang H et al., 2020b; Tang JL et al., 2021; Frühstück et al., 2022) or allowing unlimited time (Wang Y et al., 2018; Park et al., 2019; Qiao et al., 2019b; Zhou S et al., 2019; Yang H et al., 2020). Zhou S et al. (2019) evaluated models by analyzing the shortest time required for participants to differentiate between real and fake images. A smaller time threshold indicates easier distinction, implying poorer model performance. When evaluating different models, the lower bound on the time required for users to make judgments may vary. It is reasonable to speculate that if the display time of an image is smaller than its threshold, user misjudgment may be caused because the quality of the generated image is close to that of the real image.

Evaluation criteria: the evaluation criteria can be affected by three key factors. (1) Evaluation methods. As mentioned in Section 4.1, the user evaluation criteria for absolute and comparative evaluations are inconsistent. Absolute evaluation relies on the experience of users, leading to varying evaluation criteria among different users. In contrast, the user evaluation criteria for comparative evaluation depend on the shared baseline models. (2) Description of evaluation content. We introduced some inconsistencies in the description of evaluation content in Section 3.1 and recommended to adopt

a unified and clear description of evaluation content. (3) Evaluation aspects. Evaluation aspects can directly impact the results. Section 2 describes in detail the evaluation aspects used for each image generation task. Instead of using a single aspect to evaluate the performance of a generative model, we recommend using multiple evaluation aspects to obtain comprehensive results of image generation.

Some other factors may also affect experimental results. These include the workload of a single participant (Otani et al., 2023), evaluation interface/task presentation, problem formulation (Otani et al., 2023; Ioannou and Maddock, 2024), number of samples (Frolov et al., 2021; Otani et al., 2023), number of models (Frolov et al., 2021; Ioannou and Maddock, 2024), content of the final report (Frolov et al., 2021), image quality (Zhai and Min, 2020) (such as image size and image resolution), methods of generating and using images (Ioannou and Maddock, 2024) (such as displaying all results, randomly sampling some results, and displaying results in groups), annotation quality (Otani et al., 2023), compensation, and qualifications (Otani et al., 2023). Note that human evaluation involves collecting quantitative data rather than qualitative observations (Ioannou and Maddock, 2024). Similar to quantitative assessments with automated ones, reproducibility and repeatability are key factors in evaluating the reliability of human evaluation (Ioannou and Maddock, 2024). Therefore, researchers should consider and control the possible impacts of these factors when designing experiments to obtain reliable conclusions.

4.4 Data analysis methods

Data analysis in human evaluation commonly involves evaluation of annotation quality and comparison of model performances.

Annotation quality is an important aspect of image generation evaluation, and is usually measured via consistency test, which is also known as inter-annotator agreement. This test indicates the degree of agreement between the evaluation results of evaluators, and is often used to measure the disagreement between different evaluators. The higher the consistency of the evaluators' results is, the more credible the results will be (Otani et al., 2023; Ku et al., 2024). Consistency metrics (Otani et al., 2023; Ku et al., 2024), such as Cronbach's alpha, Cohen's Kappa, Fleiss' Kappa, Krippendorff's alpha, and percentage agreement, with their values ranging from 0 to 1, are commonly used.

Mean opinion score (MOS) is a popular score to measure the performance of each model. It represents the average opinion of multiple participants on the image quality, providing a uniform measure. MOS requires participants to rate images, and the scores are usually within a certain range such as 1–5, which correspond to "Bad," "Poor," "Fair," "Good," and "Excellent," respectively (Gao YX et al., 2022; Zhang KH et al., 2022; Koley et al., 2023). After each participant scores each sample, the average of all scores is calculated to obtain the MOS of the sample (Wang ZH et al., 2021):

$$\mathrm{MOS}=\frac{1}{N}\sum_{i=1}^{N}R_i, \tag{1}$$

where N is the number of evaluators involved in the scoring and R_i is the score given by the i^{th} evaluator to the test object (such as an image). A higher MOS indicates better perceived image quality. As MOS values are affected by the overall sample, the statistics of opinion scores are usually considered (Zhang KH et al., 2022). Statistical processing is necessary to ensure the accuracy and representativeness of MOS, including sample representativeness, data processing, and outlier handling.

Directly obtaining MOS usually requires organizing a large number of user studies, which may be time- and resource-consuming. The MOS prediction model provides an easy-to-use and efficient way to automatically predict MOS (Gao YX et al., 2022), which can save labor and time. We summarize some data analysis methods for consistency between the prediction model and true MOS (Wang J et al., 2023; Zhao K et al., 2023), namely, the Spearman rank correlation coefficient, Pearson linear correlation coefficient, Kendall rank correlation coefficient, and root mean squared error.

MOS is suitable for the data analysis of scale evaluation. Ranking evaluation mostly counts the percentage of each image generation model ranked at a certain level. For example, Ho TT et al. (2020) counted the percentage of each model ranked first to fifth. In addition to statistical percentages, comparative evaluations often calculate fooling rates. For example, some studies (Ma et al., 2017; Chen X et al., 2019; Zhu Z et al., 2019) used two fooling rate metrics: R2G (means the ratio of the number of real

images rated as generated to the number of real images) and G2R (means the ratio of the number of generated images rated as real images to the number of generated images).

To further reveal the performance differences between various image generation models, some studies used significance tests. For instance, Zhou S et al. (2019) used a one-way ANOVA with Tukey's pairwise post-hoc tests to compare differences between models.

5 Automatic evaluation

Automatic evaluation plays a crucial role in image generation evaluation. Computational models are used to automatically assess the image quality for developing algorithms and models that align with human perception (Ma and Fang, 2021). Compared with human evaluation, automatic evaluation has the advantages of high efficiency, objectivity, and consistency. These benefits enable the rapid processing of large-scale evaluation datasets and ensure uniform evaluation criteria in different models and experimental conditions. Thus, researchers have a reliable basis for comparison and optimization (Otani et al., 2023; Wang J et al., 2023; Ioannou and Maddock, 2024).

Current automatic evaluation is insufficient to represent human perception (Blau and Michaeli, 2018; Wu XS et al., 2023a; Cai et al., 2024; Jayasumana et al., 2024; Liu JX et al., 2024). As the quality of generated images continues to increase, the complexity of automatic evaluation increases. To promote the innovation of automatic evaluation in image generation, the evaluation metrics and evaluation benchmarks are systematically reviewed herein, particularly focusing on the advancements made in the past five years. Specifically, we review classical evaluation metrics, compare their evolution, and introduce novel metrics specific to different tasks. We focus on innovative evaluation benchmarks and demonstrate their significance in assessing generative models.

5.1 Evaluation metrics

5.1.1 Classic evaluation metrics

In the field of computer vision, various metrics have been developed to evaluate model performance; these include pixel-level metrics (PSNR and SSIM), distribution-level metrics (IS and FID), classification-oriented metrics (accuracy), perception (feature)-level metrics (LPIPS), and detection-oriented metrics (precision, recall, IoU, and R-Precision).

PSNR (Jayant and Noll, 1984) is widely used to measure image quality; however, it has notable limitations, particularly in complex scenes and when handling intricate details (Ledig et al., 2017; Blau and Michaeli, 2018). High PSNR scores do not always correspond to visually satisfactory results.

SSIM, introduced to address the limitations of PSNR, evaluates image similarity by considering luminance, contrast, and structural information (Wang Z et al., 2004). SSIM aligns better with the perception of image quality of the human visual system by focusing on local structural similarities; this makes it an important evaluation metric in tasks such as sketch-to-image generation and pose-guided image generation (Siarohin et al., 2018; Zhu Z et al., 2019; Ho TT et al., 2020; Xia et al., 2021a). Several improvements to SSIM have been proposed in subsequent research since SSIM was proposed. Multiscale structural similarity (MS-SSIM) (Wang Z et al., 2003) extends the original metric by incorporating multiscale information (Odena et al., 2017; Sushko et al., 2021). Mask-SSIM (Ma et al., 2017) introduces pose masks into the synthesized and target images, focusing specifically on the overall quality of human appearance (Siarohin et al., 2018; Zhu Z et al., 2019; Tang H et al., 2020b). Similarly, part-based SSIM (Shi et al., 2022) considers human joint structures by dividing images into multiple body part regions for calculation.

SSIM has certain limitations. Blau and Michaeli (2018) highlighted that SSIM struggles to capture high-level semantic information and complex visual features, leading to a perceptual gap with the human visual system. The inception score (Salimans et al., 2016) that uses a pretrained Inception v3 network calculates the Kullback–Leibler (KL) divergence between the class distribution of generated images and marginal distribution; it thus evaluates realism and diversity simultaneously. Inception score (IS) serves as a key metric for evaluating the quality of the generated image (Zhao B et al., 2019; Zhu Z et al., 2019; Sylvain et al., 2021; Zhang H et al., 2021; Hinz et al., 2022; Sauer et al., 2022). However, KL divergence

only captures the diversity and distribution of generated images, neglecting differences in the feature space between the generated and real images (Hinz et al., 2022). Additionally, it does not adequately address mode collapse (Borji, 2019; Hinz et al., 2022). Ma et al. (2017) introduced mask-IS to calculate the IS similar to mask-SSIM.

FID (Heusel et al., 2017) is a key metric for assessing the realism, diversity, and overall quality of generated images. It extracts feature vectors from the real and generated images using a pretrained Inception v3 model. These vectors are then modeled as multivariate Gaussian distributions, and their difference is measured with FID. FID is used to evaluate image quality using mean and covariance matrices, reflecting high-level statistical properties and aligning with human visual perception. Thus, it is suitable for complex scenes with multiple objects (Hua et al., 2021; Saseendran et al., 2021; Xia et al., 2021b; Chen P et al., 2022; Hinz et al., 2022; Cheng et al., 2023; Wu ZB et al., 2023; Zhao YQ et al., 2023). However, FID has limitations; for example, it relies on pretrained models, which may reduce generalization to new samples and make it vulnerable to manipulation (Sajjadi et al., 2018; Hinz et al., 2022; Kynkäänniemi et al., 2023). It also assumes Gaussian feature distributions (Hinz et al., 2022; Kynkäänniemi et al., 2023) and is sensitive to small sample sizes (Bińkowski et al., 2021); however, it is robust to image corruption (Sajjadi et al., 2018). Kernel inception distance (KID) (Bińkowski et al., 2021) addresses these issues using kernel methods to compute the maximum mean discrepancy (MMD) between feature distributions of generated and real images. It provides an unbiased estimate and stability for small sample sizes, making it suitable for few-shot generation tasks (He YT et al., 2023; Zhao YQ et al., 2023). SceneFID (Sylvain et al., 2021) computes FID for individual objects, allowing for the finer assessment of semantic consistency in multiobject images, particularly in complex scene generation (Li ZJ et al., 2021; Cheng et al., 2023). To better align with human evaluations, CLIP-MMD (Phung et al., 2024) uses a larger CLIP model to extract image embeddings and calculates MMD between their distributions. It thus accurately reflects quality changes in images with complex distortions.

Accuracy is a widely used objective metric; its definition and calculation methods vary across tasks. Several novel accuracy-type metrics have recently emerged for conditional image generation. Siarohin et al. (2018) and Zhu Z et al. (2019) used a pretrained residual network (ResNet) to classify the generated images based on the real images, allowing classification accuracy to objectively assess realism. Conversely, Ravuri and Vinyals (2019) introduced a classification accuracy score (CAS) to evaluate the accuracy of predicting real images using a model pretrained on synthetic images, addressing the diversity of generated outputs (Li ZJ et al., 2021). These accuracy metrics are commonly applied to conditional image generation tasks (Odena et al., 2017; Ravuri and Vinyals, 2019; Zhao B et al., 2019; Luo et al., 2020; Li ZJ et al., 2021; Sun and Wu, 2021; Sylvain et al., 2021; Cheng et al., 2023; Hassan et al., 2023). Average count accuracy (Saseendran et al., 2021) was designed for the multiclass multi-instance count conditioned image generation task, wherein instance count information is incorporated. Pixel accuracy measures the proportion of correctly classified pixels between generated images and semantic maps to evaluate semantic consistency (Park et al., 2019; Tang H et al., 2020a; Zhu Z et al., 2020; Tan et al., 2021; Chen P et al., 2022; Tang H et al., 2023). Semantic object accuracy (Hinz et al., 2022) is crucial for text-to-image generation (Zhang H et al., 2021; Yan et al., 2022), and verifies whether generated images contain the described objects. These accuracy variants considerably enhance the assessments of semantic consistency, recognizability, and fine-grained image quality.

LPIPS (Zhang R et al., 2018) further considers perceptual aspects using pretrained models to extract deep features from images. It calculates the weighted L2 distance between the corresponding feature maps of two images across various network layers using learned weights. A lower LPIPS score indicates a higher similarity in the deep features of images. It focuses on the visual similarity of images as perceived by the human eye, providing a more accurate reflection of the perceptual evaluation of image quality (Huh et al., 2020; Lu et al., 2022; Zhang PZ et al., 2022; Wu ZB et al., 2023). Intracluster pairwise LPIPS was proposed by Ojha et al. (2021), who calculated LPIPS after clustering generated images. Thus, it measures whether the model can generate visually diverse images in case of limited training samples (Ojha et al., 2021; Zhao YQ et al.,

2022). Mask-LPIPS was introduced by Ma et al. (2017) to evaluate the overall quality of foreground human image generation (Lu et al., 2022).

Metrics such as precision, recall, IoU, and R-Precision are commonly used in detection and recognition tasks as well as image generation tasks (Johnson et al., 2018; Sajjadi et al., 2018; Xu T et al., 2018; Qiao et al., 2019a; Ramesh et al., 2022; Sauer et al., 2022; Wu JY et al., 2022; He YT et al., 2023; Zhu JY et al., 2024). These metrics are also used to evaluate the recognizability of objects in generated images and semantic consistency between images and specified conditions (Kynkäänniemi et al., 2019; Sushko et al., 2021).

5.1.2 Recent evaluation metrics

We have identified several novel metrics that fully consider the characteristics of specific image generation tasks. These metrics are discussed below according to the types of image generation tasks.

In sketch-to-image generation tasks, Sketchy-COCO (Gao CY et al., 2020) introduces shape similarity to measure the semantic consistency between the generated image and sketch condition information. Koley et al. (2023) introduced the fine-grained metric (FGM) to calculate the cosine similarity between generated images and conditional sketches using features from a pretrained fine-grained sketch retrieval model.

In text-to-image generation tasks, the CLIP score (Hessel et al., 2022) uses the cross-modal representation ability of the pretrained CLIP model to extract the embedding vectors of text descriptions and generated images. The cosine similarity is calculated between two modalities, which is highly consistent with human evaluations (Chen JS et al., 2024; Li H et al., 2024; Li SK et al., 2024; Liang YW et al., 2024; Phung et al., 2024). Hall et al. (2024) introduced the object-CLIP score to assess whether generated images match prompt components and semantic consistency in complex scenes. Yan et al. (2022) proposed the spatial semantic CLIP score, which divides the generated image into five patches by focusing on the controllability and combination capability of the generative model. It thus measures the semantic consistency and spatial alignment ability. As human evaluation is expensive, several metrics have been proposed to simulate human preference evaluations. Xu JZ et al. (2023) proposed ImageReward, the first general human preference reward model for text-to-image generation. It uses the bootstrapping language-image pre-training (BLIP) model as the backbone to train a reward model on a large-scale expert comparison dataset. ImageReward can serve as an evaluation metric (Zheng WD et al., 2024) and a tool for model optimization using the ReFL method. Wu XS et al. (2023b) fine-tuned the CLIP model on the HPD v2 dataset to determine the alignment between generated images and given text prompts and assess alignment with human preference scores.

Naeem et al. (2020) proposed two metrics for few-shot image generation tasks: density and coverage. Unlike precision, density measures the expected probability that the generated fake samples appear in the dense area of real samples; i.e., how many real sample neighborhoods contain fake samples. Coverage measures the proportion of generated fake samples covering the diversity of real samples. Compared with recall, coverage quantifies this proportion by constructing a neighborhood manifold around the real samples. Density and coverage have certain applications in the study of few-shot image generation tasks (Naeem et al., 2020; Mondal et al., 2023).

In layout-to-image generation tasks, the YOLO score (Li ZJ et al., 2021) employs a pretrained YOLO model to perform object detection on the generated images, and AP, AP50, and AP75 scores are calculated to evaluate the semantic consistency between generated images and layout maps (Cheng et al., 2023). To address the lack of quality evaluation methods for boundary generation, Quan and Lang (2024) proposed BoundaryFID, which considers the boundary regions of overlapping objects and improves the handling of boundary generation issues.

In scene graph-to-image generation tasks, Tripathi et al. (2019b) proposed a relation score and a mean opinion relation score. The former focuses on geometric or spatial relationships, whereas the latter addresses semantic and nonspatial relationships. Miyake et al. (2024) proposed the positional relation of three objects (PTO) and area of overlapping (AoO). PTO evaluates the correct proportion of three bounding boxes generated by each edge in the scene graph, and AoO checks whether the introduction of ternary hyperedges can control the degree of object overlap. They help the model understand and quantify the degree of overlap and positional relationship of the generated image. Based on

GPT4V, Shen et al. (2024) proposed SceneGraph-IoU, Entity-IoU, and Relation-IoU to fully capture the relationships in complex images. These metrics are used to measure the consistency of the generated image with the scene graph, fidelity of entity generation, and accuracy of entity relationships, separately. These metrics are also used to examine the degree of integration between the generated image and scene graph conditional information in a finer manner.

Chen P et al. (2022) proposed single-image FID (SIFID) to evaluate single image generation models by capturing finer-grained differences in local details. Zhu Z et al. (2020) proposed mean class-specific diversity and mean other-classes diversity to evaluate the diversity within specific semantic regions and the degree of change in nontarget semantic regions, respectively. These metrics were proposed to ensure that other regions remained stable when modifying specific semantic regions. Tan et al. (2021) extended these metrics to the instance level.

In pose-guided image generation tasks, MPD (Chan et al., 2019) measures the difference between predicted and actual key points. It thus effectively assesses pose restoration accuracy in the generated images (Zhang PZ et al., 2022). Zhang PZ et al. (2022) used the FastReID platform to conduct pedestrian reidentification experiments. They calculated the rank-k and mean average precision metrics for synthetic image reidentification to verify the semantic consistency. Bau et al. (2019) used FSD to evaluate the difference between generated and real images in terms of object segmentation statistics to measure the perceptual distance from the generated images to real ones.

In image-to-panorama generation tasks, sharpness difference (Regmi and Borji, 2018) was proposed to measure the loss of sharpness during image generation. It is widely used to evaluate the detail preservation of generated panoramic images (Tang H et al., 2019; Wu SS et al., 2023). Omnidirectional stitching image quality assessment was proposed by Li J et al. (2019) to focus on details such as geometric deformation, chromatic aberration, and blind spots during the synthesis process. OS-IQA utilizes a linear regression classifier to match multiple local and global evaluation metrics with human subjective evaluation. It thus provides a comprehensive evaluation of the quality of stitching regions and improves consistency with human evaluation. FAED was proposed by Oh et al. (2022) to evaluate the perceptual quality for RGB-D panorama synthesis. FAED employs an auto-encoder to reconstruct inputs from latent features in an unlabeled dataset, and calculates the FID between synthetic and real RGB-D data using the learned feature distribution. Moreover, multidiffusion (Bar-Tal et al., 2023) uses FID to measure the distance between the distribution of random sampled image blocks from the panorama and the distribution of images generated by the reference model.

In class-conditional image generation tasks, Benny et al. (2021) extended the two classic unconditional metrics, namely IS and FID, by proposing conditional IS (CIS) and conditional FID (CFID). Specifically, intra-class CIS and CFID measure the quality and diversity of each conditional class in generated images, whereas interclass CIS and CFID measure the closeness between class representations in the generated data and those in the real data distribution. CIS and CFID more comprehensively reflect the performance of the conditional image generation model, and can effectively measure the overall quality and diversity of generated images.

5.2 Evaluation benchmarks

A comprehensive investigation of recent studies on evaluation benchmarks revealed a gradual shift toward using previously unseen data during training and the development of tailored evaluation metrics, as well as growing emphasis on human evaluation. In this subsection, we discuss benchmarks that focus on evaluation, which differ from the traditional benchmarks, by incorporating richer novel data and more comprehensive evaluation metrics. These evaluation benchmarks are not limited to automatic evaluation, but are designed for long-term use. As a result, they enable more comprehensive performance comparisons and validations of newly proposed methods over time. We summarize the innovations of evaluation benchmarks into three aspects: (1) a more comprehensive focus on evaluations, including the utilization of automatic and human evaluations (Bakr et al., 2023; Cho J et al., 2023; Lee et al., 2023) and improvements in automatic evaluation platforms (Lee et al., 2023); (2) an expanded focus on the aspects or capabilities of methods that have not been evaluated, such as carefully designed prompt

categories (Chen WH et al., 2022; Petsiuk et al., 2022; Saharia et al., 2022; Yu JH et al., 2022) and introducing the evaluations of aesthetics, ethics, and other aspects (Cho J et al., 2023; Lee et al., 2023); (3) an increased emphasis on exploring new automatic methods for evaluation such as introducing visual question-answering (VQA) systems (Hu et al., 2023).

In text-to-image generation, Cho J et al. (2023) proposed PaintSkills to evaluate three types of visual reasoning abilities: object recognition, object counting, and spatial relationship understanding. PaintSkills addresses social bias and provides detailed evaluation designs for issues such as gender and skin tone. By using carefully designed prompt templates about objects, quantities, and spatial relationships, it employs corresponding objective evaluation metrics and subjective human evaluations. Experimental results reveal gaps in object counting and spatial relationship understanding in text-to-image generation models (Cho J et al., 2023).

Saharia et al. (2022) proposed the equally systematic DrawBench to directly evaluate and compare the performance of different text-to-image generation models. DrawBench covers 11 prompt categories and evaluates aspects such as model fidelity, object quantity, and spatial relationships. To assess the performance differences in alignment and fidelity between two generative models, human evaluators are presented with pairs of images generated using these models based on DrawBench prompts. The evaluators are instructed to express their preferences with each pair and provide a comparative analysis of the generative capabilities of models.

PartiPrompts (Yu JH et al., 2022) emphasizes human evaluation such as DrawBench. It distinguishes itself by associating prompts with categories and challenges as separate labels, providing a more comprehensive assessment framework for image generation models. EntityDrawBench (Chen WH et al., 2022) was proposed to cover various detailed entity types. Text-image pairs were constructed for these objects, ranging from frequent to rare, providing valuable insights into the ability of generative models to handle diverse and uncommon entities. Fig. 7 presents the details of Paintskills.

Petsiuk et al. (2022) further proposed the multitask benchmark, which contains 32 types of tasks. These tasks are subdivided into automatic evaluation or human evaluation and prompt difficulty is divided into three levels from easy to hard. The corresponding images generated for different tasks at varying difficulty levels are scored to judge the performance of generative models. In addition, TISE (Dinh et al., 2022) focuses on metrics for automatic evaluation and introduces a series of new metrics to automatically evaluate aspects such as position alignment, counting, and fidelity; these aspects have been highlighted in the previously mentioned benchmarks.

HRS-Bench (Bakr et al., 2023) addresses the limitations of existing benchmarks by covering 50 diverse application scenarios and evaluating 13 novel skills in five categories: accuracy, robustness, generalization, fairness, and bias. Prompts are divided into three difficulty levels. HRS-Bench is one of the few benchmarks similar to PaintSkills, which integrates human and automatic evaluations. The high consistency between automatic and human evaluation results demonstrates its comprehensiveness and reliability in assessing image generation tasks. Fig. 8 presents the details of HRS-Bench.

The TIFA benchmark (Hu et al., 2023) incorporates a VQA system to assess image generation. It uses a pretrained large language model (LLM) to generate multiple question–answer pairs. After parsing and filtering for high-quality pairs via the question–answer system, the latest VQA model is used to assess the alignment of generated images with these pairs. Consequently, the fidelity of generated images is evaluated. Results indicate that the image quality assessments of the TIFA benchmark highly correlate with human evaluations. This also indicates that generative models still struggle with multiobject generation, spatial relationships, and abstract concepts, consistent with the findings from other benchmarks such as PaintSkills (Cho J et al., 2023). Fig. 9 presents the details of TIFA.

Davidsonian scene graph (DSG) (Cho J et al., 2024) is an improvement over traditional VQA methods such as TIFA. DSG uses a directed acyclic graph for semantic representation, where nodes denote the atomic propositions (the smallest possible semantic units) and edges denote the dependencies between them. Using this structured representation, a series of reliability problems such as duplication, ambiguity, and invalidity in traditional VQA methods can be effectively reduced. DSG covers four fine-grained evaluation aspects: entities,

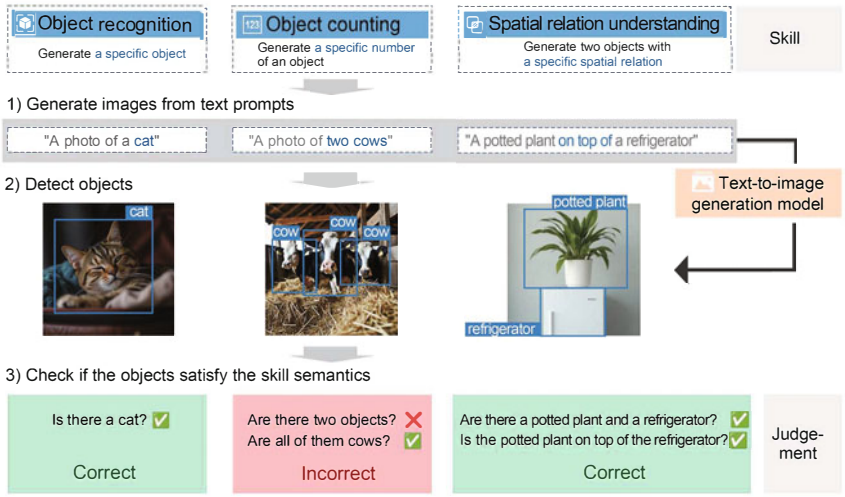

Fig. 7 Demonstration of PaintSkills (adapted from Cho J et al. (2023)). This method generates images from text prompts that require the model to possess three different visual reasoning abilities. The visual reasoning ability of the model is evaluated using object detection results and whether the generated images align with the input text prompts is determined

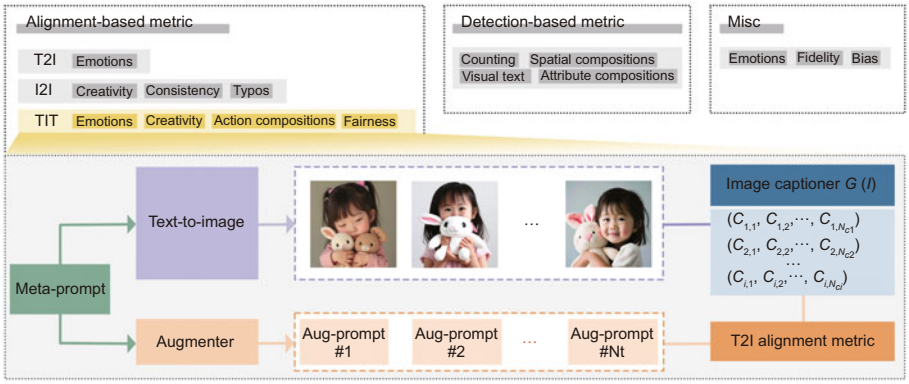

Fig. 8 Demonstration of HRS-Bench (adapted from Bakr et al. (2023)). The image on the left shows the taxonomy of this method, whereas that on the right shows the text-to-image alignment metric based on the augmented captioner. This metric analyzes the generated images using the augmented captioner to assess the alignment between the images and input text

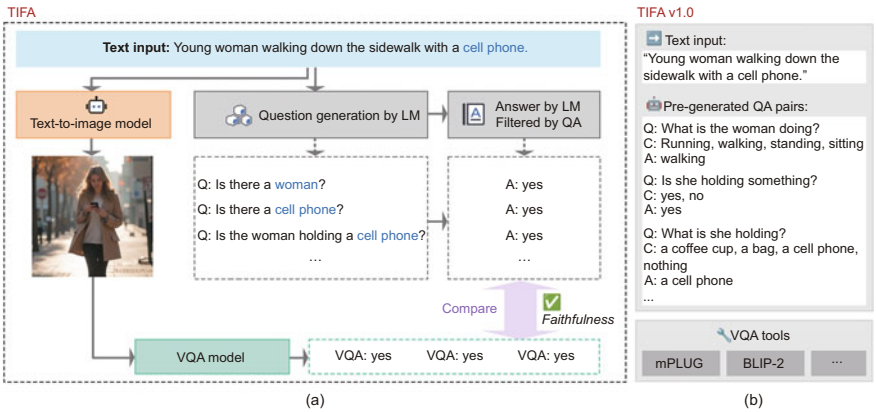

Fig. 9 Demonstration of TIFA (adapted from Hu et al. (2023)): (a) overview of TIFA for evaluating the fidelity of synthetic images; (b) TIFA v1.0 benchmark

attributes, relationships, and global context. It also collects DSG-1k, a fine-grained human-annotated benchmark with a diverse set of 1060 prompts, which can further advance the research on alignment in text-to-image generation. Fig. 10 presents the details of DSG.

TIFA (Hu et al., 2023) and DSG (Cho J et al., 2024) are divide-and-conquer methods that break text into simpler question pairs. Lin ZQ et al. (2025) proposed GenAI-bench (Fig. 11), an end-to-end method that produces an alignment score by computing the probability of a "Yes" answer to a simple "Does this figure show text?" question. Their results highlight the advantages of end-to-end methods, and indicate that divide-and-conquer methods are more complex and struggle with compositional texts. Moreover, GenAI-bench has made 15 810 human annotations publicly available, which can effectively support the development of metrics related to consistency with human evaluations.

The HEIM benchmark (Lee et al., 2023) emphasizes current ethical considerations in image generation. It addresses fundamental aspects such as alignment and quality, as well as social and ethical issues such as bias, fairness, and toxicity. In the evaluation platform, it introduces new metrics for evaluating diverse skills such as aesthetics, incorporating

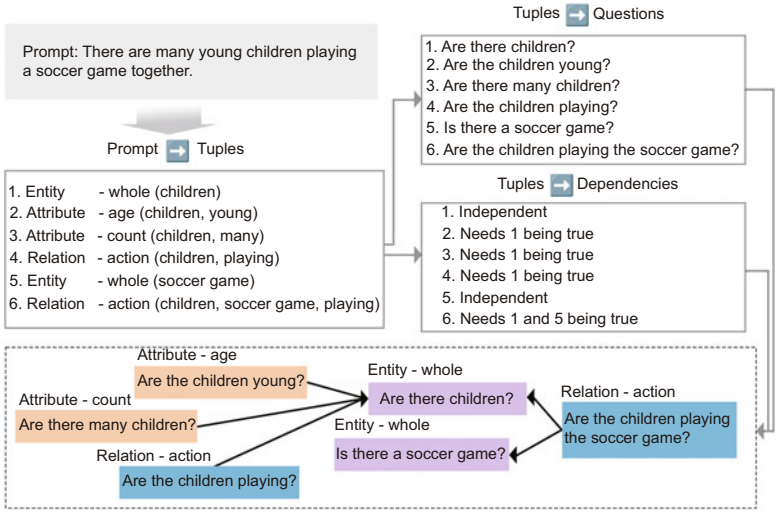

Fig. 10 Demonstration of the Davidsonian scene graph (adapted from Cho J et al. (2024)). This method first generates semantic tuples from text prompts and then converts them into a structured form

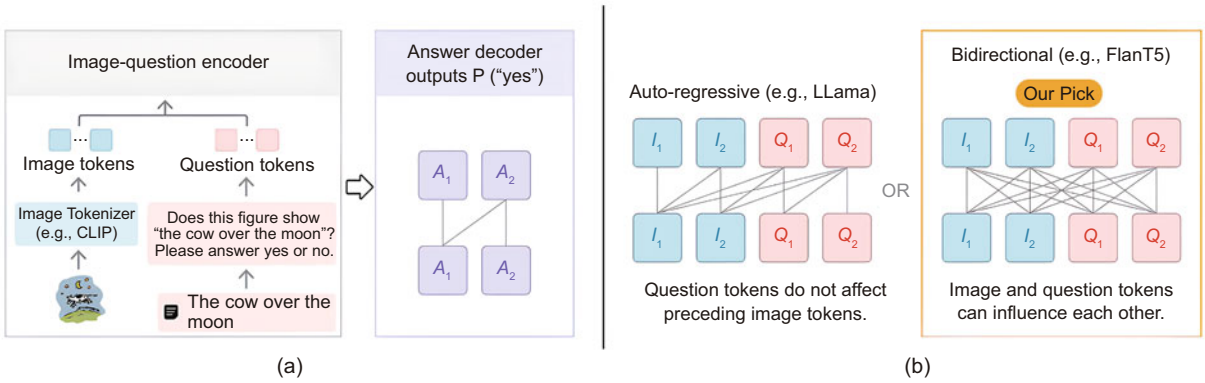

Fig. 11 Demonstration of GenAI-bench (Lin ZQ et al., 2025): (a) VQAScore evaluates the alignment between an image and text by converting the text into the question "Does this figure show 'text'? Please answer yes or no." It then inputs the image and the question into an image-question encoder and outputs the probability of "yes." (b) When selecting the image-question encoder architecture, VQAScore uses a bidirectional encoder and allows the image and question to influence each other. It thus improves the alignment between the image and text, surpassing traditional autoregressive architectures

those specific metrics for various aspects and human evaluations for ethical aspects such as bias, fairness, and toxicity. This benchmark enables a comprehensive assessment of the technical performance and social impacts of generative models.

In image-to-image generation, EditBench (Wang S et al., 2023) is a hand-curated benchmark for text-guided image inpainting based on 240 images. It was designed to capture aspects of varying categories and difficulty levels. For each image and mask pair, three different text prompt types are provided: mask simple, mask rich, and full; these prompts represent different editing specifications. EditBench evaluates models across 11 fine-grained aspects in three dimensions: attributes, objects, and scenes. It also includes two evaluation methods, namely single-image evaluations and side-by-side evaluations, which help assess model performance on fine-grained tasks and enable intuitive comparisons between models.

The Emu Edit benchmark (Sheynin et al., 2024) is a diverse and challenging image editing benchmark designed to evaluate instruction-driven image editing models. It builds a comprehensive dataset that covers three main categories, namely region editing, free-form editing, and visual tasks, comprising 16 tasks and ten million samples. The benchmark defines seven types of image editing operations and collects high-quality, challenging editing instructions via crowdsourcing. The evaluation criteria primarily focus on edit text alignment and image faithfulness. The Emu Edit benchmark places particular emphasis on the diversity of image editing and task complexity. It also effectively evaluates the accuracy and instruction execution effect of the model when performing various editing operations.

Huang KY et al. (2023) emphasized the compositional abilities of generative models by proposing T2I-CompBench and introduced generative model finetuning with reward-driven sample selection (GORS) for finetuning. Huang ST et al. (2024) innovatively presented ActionBench in text-to-image generation with additional action condition information to evaluate the capabilities of generative models in action customization. Meng et al. (2024) evaluated the physical commonsense of generative models and validated the effectiveness of the benchmark using scores from GPT-4o and human evaluation.

Although comprehensive and detailed evaluation benchmarks have been widely proposed for text-to-image generation and image-to-image generation, similar efforts have been rarely employed in other image generation tasks. Table 4 compares benchmarks used in text-to-image generation evaluation. Benchmarks for other tasks often lack thorough

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

Specified skill (SS) is the modeling capability that the benchmark is designed to evaluate. For example, PaintSkills evaluates five abilities: object recognition, object counting, spatial relationship understanding, gender bias, and skin tone bias. Although TIFA 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 the 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

consideration of datasets, metrics, and implementation details; therefore, they fall short of the defined ideal evaluation benchmarks. For example, Zhou S et al. (2019) introduced the HYPE benchmark, which combines psychophysics for human evaluation of generative models. However, its data and metrics still need the comprehensive design. In other image generation tasks, most evaluation studies have primarily focused on either datasets or metrics and provided innovations related to specific issues in datasets or metrics (Naeem et al., 2020; Benny et al., 2021; Hinz et al., 2022; Jayasumana et al., 2024; Quan and Lang, 2024).

6 Challenges and future directions

6.1 Challenges

Herein, we discuss the challenges in image generation evaluation in two parts: evaluation protocols and evaluation methods.

6.1.1 Evaluation protocols

Existing studies have not yet reached a unified evaluation protocol. We describe the challenges in determining the evaluation aspects and datasets.

1. Evaluation aspects

The challenges can be divided into two parts:

(1) Lack of generally accepted guidelines for selecting evaluation aspects: evaluation aspects directly impact the output; however, there is currently no unified guideline for selecting these aspects. The evaluation aspects adopted by the existing studies may not be sufficiently comprehensive, and the evaluation content may be limited.

First, only automatic evaluation is used. As discussed in Section 3.2, some image generation tasks are evaluated using only automatic metrics. However, these metrics do not always align with human evaluations, leading to potentially inaccurate results (Wang Z et al., 2003; Lin TY et al., 2014; Shocher et al., 2020; Petsiuk et al., 2022; Ku et al., 2024). Second, only a few automatic metrics are used. For example, in few-shot image generation, the utility and generalization of generated images are crucial. However, only a few studies (Hong Y et al., 2020; Gu et al., 2021; Xie et al., 2022; Li LX et al., 2023) have considered these aspects and used accuracy to evaluate the performance of generated images in classification tasks.

(2) Lack of comprehensive and universal evaluation benchmarks: except for text-to-image and image-to-image generation, other image generation tasks lack dedicated and continuously updated evaluation benchmarks. Instead, they rely on ad hoc benchmarks with limited datasets and evaluation metrics.

2. Datasets

The challenges can be concluded in three parts:

(1) Accessibility threshold: although most large datasets are open to the public, they cannot be easily used by individual users due to the substantial hardware requirements involved (Alhabeeb and Al-Shargabi, 2024).

(2) Long-tail effect: the long-tail effect is a phenomenon in which the frequency distribution of objects of different categories is uneven in an image dataset, and most object categories have a low frequency of occurrence. This distribution presents several challenges, including data sparsity (infrequently occurring object categories in the training dataset hinder the model to learn effective representations (Cao et al., 2019)), generation inconsistency (low-frequency object categories must also produce high-quality and consistent images during generation (Zhang YK et al., 2023)), and difficulty in processing spatial relationships (complex spatial and semantic relationships between multiple objects must be handled, such as occlusion and overlap (Ashual and Wolf, 2019; Li ZJ et al., 2021)).

(3) Lack of high-quality, large-scale, and open datasets: this challenge affects nearly all image generation tasks. First, due to differences in human perception, the annotation scores of real datasets usually have high variance. This makes it difficult for evaluation models to capture human scoring rules (Wang JR et al., 2023). Second, the difficulty in evaluating the existing datasets varies (Zhou S et al., 2019), such as digital image MNIST, face datasets, and complex object datasets. Most studies curate their own datasets; therefore, different models cannot be compared (Ku et al., 2024). Additionally, large-scale datasets are usually collected by large companies and are not open to the public, which leads to the research limitations and evaluation difficulties (Kirstain et al., 2023).

6.1.2 Evaluation methods

The challenges include those in both human and automatic evaluation methods, as well as the inconsistencies between them.

1. Human evaluation

We can conclude three challenges as follows:

(1) Individual differences among participants: individual differences among participants can affect the reliability and reproducibility of human evaluation results (Wang ZH et al., 2021; Gao YX et al., 2022; Ku et al., 2024). Due to varying standards and preferences, participants may provide different assessments for the same image; this can yield inconsistent results and difficulty in distinguishing performance differences between models (Zhou S et al., 2019). Moreover, accurately quantifying human evaluation remains a critical challenge.

(2) Lack of empirical research on evaluation methods: human evaluation methods are generally divided into absolute and comparative approaches, with different opinions among researchers. Otani et al. (2023) supported absolute evaluation, arguing that comparative methods are problematic due to issues such as outdated shared baseline models and limited interpretability. Similarly, Khashabi et al. (2022) suggested using absolute evaluation to enable comparisons across all previous models. In contrast, Xu QQ et al. (2012) highlighted that absolute evaluation suffers from vague scale definitions and varied user interpretations, favoring comparative evaluation instead. Li BY et al. (2019) argued that comparative evaluation is more robust and consistent in subjective experiments. However, the specific effects of these methods on evaluation outcomes remain unclear.

(3) Temporal and financial cost: human evaluation is time-consuming and costly (Zhou S et al., 2019; Frolov et al., 2021; Xu JZ et al., 2023). Researchers need to dedicate considerable effort into recruiting, screening, and training evaluators, although the process of individually assessing images is laborious. Additionally, accuracy and time of human evaluation share an inherent trade-off (Zhou S et al., 2019). In general, increasing the number of evaluators can enhance accuracy, but it also leads to higher time and financial costs.

2. Automatic evaluation

We list the challenges in three main parts, namely the scope of applicability, the reliability of evaluation results, and the interpretability of evaluation results.

For scope of applicability of automatic evaluation, the challenges are categorized into the following three types:

(1) Limited evaluation scope: Hu et al. (2023) reported that DALL-Eval was applicable only to synthetic text and not naturally generated text. Moreover, its evaluation criteria overlook many important image features such as activities, geographic locations, weather, time, materials, shapes, and sizes. The range that a single metric can evaluate is limited and cannot cover all aspects of evaluation. For instance, FID and IS are widely used to assess image quality, realism, and diversity, but neither of these metrics can be used to evaluate the alignment between the input text and generated image (Frolov et al., 2021).

(2) Lack of metrics that meet evaluation requirements: Sylvain et al. (2021) noted that IS was limited by training dataset issues, whereas FID struggled with layout-to-image generation. They proposed SceneFID for multiobject evaluation to address these issues. Wu XS et al. (2023a) highlighted that mainstream metrics (e.g., IS, FID, and CLIP) failed to capture human preferences, which is a crucial factor in image quality. As mentioned in Section 3, a few automatic metrics address user preferences, with most relying on human evaluation.

(3) Insufficient integration of domain-specific expertise in existing metrics: specialized fields such as architectural planning, interior design, product design, medical imaging, and interactive interface design have specialized domain knowledge. However, most current metrics are based on statistical learning, and not specifically designed to reflect the professional knowledge requirements for evaluating the quality of generated images.

The challenges that affect the reliability of evaluation results are categorized into evaluation models and generative models:

(1) Issues of evaluation models impacting result reliability: the reliability of metrics depends on pretrained models used for evaluation, and the inductive biases and errors of these models can influence the results. For example, IS and FID have been discredited for their application to non-ImageNet datasets (Zhou S et al., 2019; Hinz et al., 2022).

FID may have different values for images in JPG or PNG format (Parmar et al., 2022). However, it is currently unclear whether CLIP can effectively measure the quality of generated images (Wu XS et al., 2023a).

(2) Issues of generative models impacting result reliability: we have categorized three types of issues.

First, issues may occur in the model during image generation. If the evaluation model fails to promptly detect issues using the generative model, false positives may be obtained. For instance, many studies on few-shot image generation use FID. However, the results obtained using FID alone may lack accuracy. Due to limited data available for few-shot image generation, FID may not effectively identify model overfitting. Some studies (Ojha et al., 2021; Zhu JY et al., 2024) have alleviated this challenge by employing a richer dataset for reliable results. Zhu JY et al. (2024) calculated the standard deviation of their FID results over five runs, whereas Ojha et al. (2021) supplemented FID with user studies.

Second, we have identified several issues with generative models that the existing metrics may fail to detect. (1) Generative infidelity: This includes entity missing, entity leakage, and attribute leakage (Lucic et al., 2018; Borji, 2019; Kynkäänniemi et al., 2019, 2023; Bińkowski et al., 2021). (2) Generated scenes may not conform to physical laws such as lighting, shadows, and multiobject interactions (Baraheem et al., 2023; Zhou MQ et al., 2024). (3) The complex nature of semantic information complicates the evaluation of whether generation is controllable and consistent with input (Park et al., 2019; Yan et al., 2022). (4) The generation of complex scenes, including those using multiple objects and intricate details such as new traditional Chinese style paintings or large-format generation (Xu HR et al., 2023), is challenging to evaluate.

Finally, from the perspective of ethical and social risks, generative models inherently possess the potential for negative societal impacts (Wu XS et al., 2023b). Advances in generative models make it increasingly difficult to distinguish between the true and false content. This shortcoming can be misused to disseminate misinformation or harmful content, such as not-safe-for-work material (Li G et al., 2022; Wu XS et al., 2023b; Yang Y et al., 2024). Furthermore, there is a risk of amplifying existing biases and stereotypes in training data (Wu XS et al., 2023b). The evaluation of indicators such as safety, fairness, and bias used to monitor these potential issues is crucial for future application; however, detecting and aligning values in the image generation model are challenging. Further research is needed to address these issues (Lee et al., 2023).

For interpretability of evaluation results, the challenges are as follows. Current metrics lack interpretability, often provide a limited view by focusing primarily on the distribution overlap, and neglect more intuitive, visually comprehensible assessments. For instance, FID fails to capture gradual improvements in iterative text-to-image generation models. Moreover, a clear correlation is lacking between FID score changes and human perception of image quality. The numerical shifts in FID do not directly indicate improvements in quality or explain specific flaws in low-FID failure cases (Sajjadi et al., 2018; Jayasumana et al., 2024).

3. Relationships between automatic evaluation and human evaluation

Many automatic evaluation metrics have weak correlations with human evaluation metrics (Lee et al., 2023; Wang J et al., 2023; Wu XS et al., 2023a; Xu JZ et al., 2023; Ku et al., 2024). These include classic metrics such as SSIM and PSNR (Zhu Z et al., 2019; Zhang XB et al., 2023; Liu JX et al., 2024), as well as popular metrics such as IS, FID, and CLIP (Zhou S et al., 2019; Zhang H et al., 2021; Ramesh et al., 2022; Yan et al., 2022; Otani et al., 2023; Wu XS et al., 2023a). In addition to the limitations of automatic evaluations discussed in Section 6.1.2, the inherent differences between automatic and human evaluations, as well as their interconnections, can lead to inconsistent results.

(1) Inherent differences between human and automatic evaluation: humans accumulate extensive visual knowledge from birth; this allows them to intuitively assess image quality without reference information by relying solely on visual perception (Wang JR et al., 2023). In contrast, automatic evaluation relies on pretrained models, and the results depend on the quality of those models. However, current automatic metrics cannot fully capture human perception (Cai et al., 2024). Human evaluation is considered the gold standard for image assessment, particularly in areas involving subjective perception and judgment such as language understanding and realism (Zhu WH et al., 2018; Zhai and

Min, 2020; Ding et al., 2021; Ma and Fang, 2021; Lee et al., 2023).

(2) Close relationship between human and automatic evaluations: data from human evaluations are essential for training and validating automatic evaluations (Xu QQ et al., 2012; Zhai and Min, 2020). However, training automatic evaluations that closely align with human ratings is challenging (Xu QQ et al., 2012; Wang JR et al., 2023). Moreover, human evaluations are influenced by individual differences and evaluation details, resulting in high variance in outcomes. This makes it difficult for automatic evaluations to accurately capture the patterns of human ratings. Human evaluations are also generally more time-consuming and expensive than automatic evaluations (Zhou S et al., 2019; Wang ZH et al., 2021; Wang JR et al., 2023; Ku et al., 2024).

6.2 Future directions

6.2.1 Evaluation protocols

1. Evaluation aspects

The future directions toward evaluation aspects in evaluation protocols are listed as follows:

(1) Developing widely accepted guidelines for selecting evaluation aspects is a promising research direction. We also recommend referring to the protocols in Section 3 for evaluation aspects.

(2) For the lack of comprehensive and universal evaluation benchmarks, we propose several suggestions for designing new evaluation metrics. First, important aspects that have received limited attention in existing evaluation studies must be considered. Recent evaluation benchmarks for text-to-image generation (Bakr et al., 2023; Lee et al., 2023; Yang Y et al., 2024) cover new evaluation metrics such as toxicity, bias, fairness, multilinguality, and robustness. These evaluation metrics have been mostly ignored in previous studies but are crucial for practical applications. Second, conducting multidimensional evaluation is beneficial. Humans evaluate images through multidimensional cognition. Multidimensional evaluation is a recent trend in research. Betti et al. (2023), Hu et al. (2023), Xu JZ et al. (2023), Zhang F et al. (2024), and Zhang SX et al. (2024) showed that multidimensional evaluation helped improve the consistency between automatic evaluation and human evaluation results. Third, collecting diverse types of information is advisable. Liang YW et al. (2024) collected extensive human feedback, effectively improving the semantic consistency of text-to-image models. Cho J et al. (2024) enhanced the traditional VQA evaluation model with directed acyclic graphs, improving the result reliability.

2. Datasets

For the future directions in datasets, we summarize the methods for constructing image datasets, and identify the key areas. These methods can be broadly classified into three categories: manual methods, automatic annotation methods, and direct data generation methods.

(1) Manual methods, such as ImageNet (Deng et al., 2009) and CIFAR-10 (Krizhevsky et al., 2009), are traditional approaches for dataset construction. The candidate images in datasets are obtained from image search engines and manually annotated and filtered. Although manual annotation is accurate, it is challenging to scale and often limited in scope. Due to individual differences among annotators, datasets may exhibit bias (Torralba and Efros, 2011; Yao et al., 2016).

(2) Automatic annotation methods are categorized into traditional learning-based methods and LLM-based methods. Traditional learning-based methods typically involve classifiers (Li LJ et al., 2007), ranking (Schroff et al., 2011), clustering, and propagation (Hua and Li, 2015) to construct datasets. In contrast, LLM-based methods (Wang Y et al., 2024) generally use prompt engineering to leverage LLMs for more automated annotation. This process considerably increases the openness of the annotation vocabulary. Compared with manual annotation, automatic annotation reduces costs and somewhat alleviates scalability issues.

(3) Direct data generation uses generative models rather than collects direct data from the real world. It is an emerging dataset construction method (Zhou YF et al., 2024). Using this approach, an unlimited number of high-quality samples can be generated for various tasks with minimal pre-training of the generative model.

Overall, future research should prioritize leveraging LLMs to enhance automatic dataset annotation methods and refine metrics for evaluating dataset quality. Research in these areas will enhance the efficiency of dataset construction and ensure high quality and reliability in the creation of larger-scale and more open datasets.

6.2.2 Evaluation methods

1. Human evaluation

For human evaluation in evaluation methods, we propose corresponding future directions in three aspects.

First, for individual differences among participants, five promising directions are listed as follows:

(1) Developing detailed evaluation instructions. Inconsistencies in annotation details across studies may lead to user misinterpretations, ultimately affecting the accuracy and interpretability of results (Ioannou and Maddock, 2024). Specifically, when evaluating "the best image," users may base their judgments on personal understanding and prior experience; this makes it difficult to explain the exact reasons behind their evaluation scores. In contrast, when evaluating "shape consistency," the scope is clearly defined, enabling a precise correlation of scores with shape consistency rather than other factors. Therefore, finding clear and standardized user evaluation instructions is a feasible research direction. Recent studies (He YT et al., 2023; Hu et al., 2023; Lee et al., 2023; Otani et al., 2023; Wu XS et al., 2023b; Xu JZ et al., 2023; Ku et al., 2024; Liang YW et al., 2024) have already adopted this approach.

(2) Formulating standardized evaluation protocols. Standardized protocols help structure user studies and mitigate the impact of inconsistent evaluation details (Frolov et al., 2021). For instance, Otani et al. (2023) proposed a human evaluation protocol for text-to-image generation, including detailed information such as datasets and annotators. Extending standardized evaluation protocols to cover various image generation tasks is a promising research direction because the evaluation requirements vary across tasks. In Section 4.3, we discussed the inconsistencies in human evaluation details and highlighted the need for further research into their specific impact on results.

(3) Developing reliable quantitative tools for human perception. Introducing psychometrics to design standardized evaluation scales is a potential research direction. On one hand, psychometric tools have demonstrated reliability and validity. Psychometrics, as a discipline that combines psychology and statistics, is used to develop scales and other tools that allow for accurate and reliable quantitative measurements of complex human psychology and perception (DeVon, 2007; Graziotin et al., 2021). Psychometric tools have been widely and successfully used in user studies in psychological research, educational evaluation, organizational behavior, and other fields (Machajdik and Hanbury, 2010). However, employing standardized scales for the human evaluation of generated images is a feasible approach. First, scale evaluation is a typical absolute method in image generation evaluation. Second, existing studies have often used inconsistent evaluation questions and scoring systems (e.g., 4-point or 5-point scale), which may undermine scoring consistency among participants and complicate comparisons between models. Thus, the introduction of psychometrics to develop reliable human evaluation tools is worth exploring in future research.

(4) Exploiting existing standardized evaluation resources and developing new resources. Existing studies (Lee et al., 2023; Ku et al., 2024; Liang YW et al., 2024; Zhang SX et al., 2024; Lin ZQ et al., 2025) have made their datasets, benchmarks, and human annotations publicly available, contributing to the design of standardized evaluation experiments. We therefore recommend leveraging these resources and encourage researchers to develop new standardized resources to cover a broader range of evaluation tasks.

(5) Annotation quality management. We recommend implementing proper quality management throughout the human evaluation process. This includes selecting and training annotators, creating and improving annotation protocols and guidelines, providing detailed annotator agreements, ensuring correct data selection, validating data, and estimating error rates (Klie et al., 2024). When disagreements occur in annotation results, simply excluding the minority noise is not enough and the reasons for discrepancies should be carefully analyzed (Fleisig et al., 2024). Human evaluation process is iterative, e.g., removing participants who fail in the qualification tests, and recruiting replacements or updating annotation protocols. We also encourage researchers to publicly share all annotation details; this transparency improves the credibility of results and enhances comparability across different studies.

Second, with regard to the lack of empirical research on evaluation methods, empirical research

into evaluation methods is essential to determine the most suitable approach.

Third, for temporal and financial cost, achieving a delicate but generalizable balance among accuracy, time, and cost remains a research direction to be explored.

2. Automatic evaluation

For automatic evaluation in evaluation methods, we provide the following future directions:

(1) Scope of applicability: first, the advantage of using multiple metrics should be emphasized to obtain comprehensive and accurate results for the limited evaluation scope. Second, to address the lack of metrics that meet evaluation requirements, improving existing metrics and developing new metrics remain promising research directions. Additionally, assessing whether pretrained models are suitable for evaluating the generated data in the current task before use is worth emphasizing. Third, developing evaluation metrics tailored to the specialized fields represents a promising research direction for addressing insufficient integration of domain-specific expertise in existing metrics.

(2) Reliability of evaluation results: we propose three promising research directions for issues of evaluation models impacting result reliability. First, the pretrained models on which the reliability of metrics depends must be updated in a timely manner as the models undergo developments. Second, general and format-insensitive evaluation methods should be developed. Third, multidimensional comprehensive evaluation platforms could be developed. For issues of generative models impacting result reliability, we highlight that developing evaluation models that specifically address the problems inherent in generative models is a promising research direction, and we suggest that subsequent evaluation models need to pay attention to a series of problems that may arise in the generative model.

(3) Interpretability of evaluation results: collecting large amounts of fine-grained data and diverse data types enhances the interpretability of evaluation results. Compared with a single total score, multidimensional data have a higher interpretability. Hu et al. (2023) proposed TIFA, which was the first to achieve fine-grained evaluation of semantic consistency for text-to-image generation via the VQA system. TIFA comprehensively evaluated the model, covering 12 aspects: objects, animals/humans, attributes, activities, spatial relationships, locations, colors, counting, food, materials, shapes, and others. Furthermore, compared with the previous automatic metrics, TIFA showed a higher correlation with human evaluations. Zhang SX et al. (2024) proposed the multidimensional preference score by learning and leveraging multidimensional human preferences. This metric captures the performance of text-to-image models in four dimensions: aesthetics, semantic alignment, detail quality, and overall assessment. In addition, collecting information related to scores facilitates more intuitive interpretation of results and improved models. Liang YW et al. (2024) introduced rich automatic human feedback, which gathers extensive human feedback, including marking implausible/misaligned image regions, annotating misrepresented or missing words in the text prompt, and providing four fine-grained scores for image plausibility, text-image alignment, aesthetics, and overall rating. This feedback constructs an interpretable and attributable evaluation model, thereby contributing to a better alignment between generated models and text.

3. Differences between automatic evaluation and human evaluation

For differences between automatic evaluation and human evaluation in evaluation methods, we provide the following two directions:

(1) For the inconsistency between automatic evaluation and human evaluation, combining automatic evaluation with user studies is the future trend.

(2) For the close relationship between human and automatic evaluation, developing generalized automatic models that approximate human evaluations remains a key research focus.

7 Conclusions

This paper presents a state-of-the-art review of image generation evaluation, summarizing a decade of research. Image generation was categorized into 10 types based on input conditions to cover general tasks. After understanding various image generation tasks, we found that their evaluation aspects were different. To distinguish this difference, six common important evaluation aspects corresponding to the requirements of each image generation task were identified, and a new evaluation protocol was developed. Compared with automatic evaluation,

fewer studies have used human evaluation, and human evaluation aspects are relatively homogeneous. To this end, human and automatic evaluations were discussed in detail. To the best of our knowledge, this is the first systematic and comprehensive introduction to human evaluation. We suggest that more works related to human evaluation can be invested in the future. Based on the reviewed studies, new content on automatic evaluation in the past five years was summarized and supplemented, including evaluation metrics and benchmarks. Finally, the challenges and potential directions of image generation evaluation in terms of evaluation protocols and evaluation methods were summarized. In summary, this paper provides strong support for systematically understanding image generation evaluation and promoting the development of this field.

Contributors

Qi LIU and Zejian LI initialized the idea and designed the review framework. Qi LIU, Shuanglin YANG, and Zejian LI collected the literature for review and created all figures and tables. Qi LIU and Shuanglin YANG drafted the paper. Zejian LI, Lefan HOU, and Chenye MENG helped organize the paper. Ying ZHANG and Lingyun SUN reviewed the paper and checked the writing details. All the authors revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

References

Agustsson E, Tschannen M, Mentzer F, et al., 2019. Generative adversarial networks for extreme learned image compression. Proc IEEE/CVF Int Conf on Computer Vision, p.221-231. https://doi.org/10.1109/ICCV.2019.00031

Ak K, Kassim A, Lim JH, et al., 2019. Attribute manipulation generative adversarial networks for fashion images. Proc IEEE/CVF Int Conf on Computer Vision, p.10540-10549. https://doi.org/10.1109/ICCV.2019.01064

Alhabeeb SK, Al-Shargabi AA, 2024. Text-to-image synthesis with generative models: methods, datasets, performance metrics, challenges, and future direction. *IEEE Access*, 12:24412-24427. https://doi.org/10.1109/ACCESS.2024.3365043

Alqahtani H, Kavakli-Thorne M, Kumar G, 2021. Applications of generative adversarial networks (GANs): an updated review. *Arch Comput Methods Eng*, 28(2):525-552. https://doi.org/10.1007/s11831-019-09388-y

Andriluka M, Pishchulin L, Gehler P, et al., 2014. 2D human pose estimation: new benchmark and state of the art analysis. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3686-3693. https://doi.org/10.1109/CVPR.2014.471

Ashual O, Wolf L, 2019. Specifying object attributes and relations in interactive scene generation. Proc IEEE/CVF Int Conf on Computer Vision, p.4560-4568. https://doi.org/10.1109/ICCV.2019.00466

Bakr EM, Sun PZ, Shen XQ, et al., 2023. HRS-bench: holistic, reliable and scalable benchmark for text-to-image models. Proc IEEE/CVF Int Conf on Computer Vision, p.19984-19996. https://doi.org/10.1109/ICCV51070.2023.01834

Baraheem SS, Le TN, Nguyen TV, 2023. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. *Artif Intell Rev*, 56(10):10813-10865. https://doi.org/10.1007/s10462-023-10434-2

Bar-Tal O, Yariv L, Lipman Y, et al., 2023. MultiDiffusion: fusing diffusion paths for controlled image generation. Proc 40th Int Conf on Machine Learning, p.1737-1752.

Bau D, Zhu JY, Wulff J, et al., 2019. Seeing what a GAN cannot generate. Proc IEEE/CVF Int Conf on Computer Vision, p.4501-4510. https://doi.org/10.1109/ICCV.2019.00460

Benny Y, Galanti T, Benaim S, et al., 2021. Evaluation metrics for conditional image generation. *Int J Comput Vis*, 129(5):1712-1731. https://doi.org/10.1007/s11263-020-01424-w

Betti F, Staiano J, Baraldi L, et al., 2023. Let's ViCE! Mimicking human cognitive behavior in image generation evaluation. Proc 31st ACM Int Conf on Multimedia, p.9306-9312. https://doi.org/10.1145/3581783.3612706

Bińkowski M, Sutherland DJ, Arbel M, et al., 2021. Demystifying MMD GANs. https://arxiv.org/abs/1801.01401

Blau Y, Michaeli T, 2018. The perception-distortion tradeoff. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6228-6237. https://doi.org/10.1109/CVPR.2018.00652

Borji A, 2019. Pros and cons of GAN evaluation measures. *Comput Vis Image Underst*, 179:41-65. https://doi.org/10.1016/j.cviu.2018.10.009

Brock A, Donahue J, Simonyan K, 2019. Large scale GAN training for high fidelity natural image synthesis. https://arxiv.org/abs/1809.11096

Cai ZP, Mueller M, Birkl R, et al., 2024. L-MAGIC: language model assisted generation of images with coherence. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7049-7058. https://doi.org/10.1109/CVPR52733.2024.00673

Cao KD, Wei C, Gaidon A, et al., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. Proc 33rd Int Conf on Neural Information Processing Systems, p.1567-1578.

Chan C, Ginosar S, Zhou TH, et al., 2019. Everybody dance now. Proc IEEE/CVF Int Conf on Computer Vision, p.5932-5941. https://doi.org/10.1109/ICCV.2019.00603

Chang XJ, Ren PZ, Xu PF, et al., 2023. A comprehensive survey of scene graphs: generation and application. *IEEE Trans Patt Anal Mach Intell*, 45(1):1-26. https://doi.org/10.1109/TPAMI.2021.3137605

Chen JS, Ge CJ, Xie EZ, et al., 2024. PIXART-Σ: weak-to-strong training of diffusion transformer for 4k text-to-image generation. 18th European Conf on Computer Vision, p.74-91. https://doi.org/10.1007/978-3-031-73411-3_5

Chen JX, Fan JY, Ye HC, et al., 2023. Exploring kernel-based texture transfer for pose-guided person image generation. *IEEE Trans Multim*, 25:7337-7349. https://doi.org/10.1109/TMM.2022.3221351

Chen P, Li ZJ, Zhang YK, et al., 2022. USIS: a unified semantic image synthesis model trained on a single or multiple samples. *Neurocomputing*, 514:70-82. https://doi.org/10.1016/j.neucom.2022.09.092

Chen WH, Hu HX, Saharia C, et al., 2022. Re-Imagen: retrieval-augmented text-to-image generator. https://arxiv.org/abs/2209.14491

Chen X, Song J, Hilliges O, 2019. Unpaired pose guided human image generation. https://doi.org/10.48550/arXiv.1901.02284

Chen YC, Shen XH, Lin Z, et al., 2019. Semantic component decomposition for face attribute manipulation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9851-9859. https://doi.org/10.1109/CVPR.2019.01009

Cheng JX, Liang X, Shi XJ, et al., 2023. LayoutDiffuse: adapting foundational diffusion models for layout-to-image generation. https://arxiv.org/abs/2302.08908

Cho J, Zala A, Bansal M, 2023. DALL-EVAL: probing the reasoning skills and social biases of text-to-image generation models. Proc IEEE/CVF Int Conf on Computer Vision, p.3020-3031. https://doi.org/10.1109/ICCV51070.2023.00283

Cho J, Hu YS, Baldridge J, et al., 2024. Davidsonian scene graph: improving reliability in fine-grained evaluation for text-to-image generation. https://doi.org/10.48550/arXiv.2310.18235

Cho SJ, Ji SW, Hong JP, et al., 2021. Rethinking coarse-to-fine approach in single image deblurring. Proc IEEE/CVF Int Conf on Computer Vision, p.4621-4630. https://doi.org/10.1109/ICCV48922.2021.00460

Croitoru FA, Hondru V, Ionescu RT, et al., 2023. Diffusion models in vision: a survey. *IEEE Trans Patt Anal Mach Intell*, 45(9):10850-10869. https://doi.org/10.1109/TPAMI.2023.3261988

Deng J, Dong W, Socher R, et al., 2009. ImageNet: a large-scale hierarchical image database. IEEE Conf on Computer Vision and Pattern Recognition, p.248-255. https://doi.org/10.1109/CVPR.2009.5206848

DeVon HA, Block ME, Moyle-Wright P, et al., 2007. A psychometric toolbox for testing validity and reliability. *J Nurs Schol*, 39(2):155-164. https://doi.org/10.1111/j.1547-5069.2007.00161.x

Ding M, Yang ZY, Hong WY, et al., 2021. CogView: mastering text-to-image generation via Transformers. Proc 35th Int Conf on Neural Information Processing Systems, Article 1516.

Ding M, Zheng WD, Hong WY, et al., 2022. CogView2: faster and better text-to-image generation via hierarchical Transformers. Proc 36th Int Conf on Neural Information Processing Systems, Article 1229.

Dinh TM, Nguyen R, Hua BS, 2022. TISE: bag of metrics for text-to-image synthesis evaluation. 17th European Conf on Computer Vision, p.594-609. https://doi.org/10.1007/978-3-031-20059-5_34

Dong YS, Tan W, Tao DC, et al., 2022. CartoonLossGAN: learning surface and coloring of images for cartoonization. *IEEE Trans Image Process*, 31:485-498. https://doi.org/10.1109/TIP.2021.3130539

Duan YP, Han CY, Tao XM, et al., 2020. Panoramic image generation: from 2-D sketch to spherical image. *IEEE J Sel Top Signal Process*, 14(1):194-208. https://doi.org/10.1109/JSTSP.2020.2968772

Elasri M, Elharrouss O, Al-Maadeed S, et al., 2022. Image generation: a review. *Neur Process Lett*, 54(5):4609-4646. https://doi.org/10.1007/s11063-022-10777-x

Fan MH, Wang WJ, Yang WH, et al., 2020. Integrating semantic segmentation and retinex model for low-light image enhancement. Proc 28th ACM Int Conf on Multimedia, p.2317-2325. https://doi.org/10.1145/3394171.3413757

Farshad A, Yeganeh Y, Chi Y, et al., 2023. SceneGenie: scene graph guided diffusion models for image synthesis. Proc IEEE/CVF Int Conf on Computer Vision, p.88-98. https://doi.org/10.1109/ICCVW60793.2023.00016

Fleisig E, Blodgett SL, Klein D, et al., 2024. The perspectivist paradigm shift: assumptions and challenges of capturing human labels. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.2279-2292. https://doi.org/10.18653/v1/2024.naacl-long.126

Foo LG, Rahmani H, Liu J, 2023. AI-generated content (AIGC) for various data modalities: a survey. https://arxiv.org/abs/2308.14177

Frolov S, Hinz T, Raue F, et al., 2021. Adversarial text-to-image synthesis: a review. *Neur Netw*, 144:187-209. https://doi.org/10.1016/j.neunet.2021.07.019

Frühstück A, Singh KK, Shechtman E, et al., 2022. InsetGAN for full-body image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7713-7722. https://doi.org/10.1109/CVPR52688.2022.00757

Gao CY, Liu Q, Xu Q, et al., 2020. SketchyCOCO: image generation from freehand scene sketches. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5173-5182. https://doi.org/10.1109/CVPR42600.2020.00522

Gao YX, Min XK, Zhu YC, et al., 2022. Image quality assessment: from mean opinion score to opinion score distribution. Proc 30th ACM Int Conf on Multimedia, p.997-1005. https://doi.org/10.1145/3503161.3547872

Graziotin D, Lenberg P, Feldt R, et al., 2021. Psychometrics in behavioral software engineering: a methodological introduction with guidelines. *ACM Trans Softw Eng Methodol*, 31(1):7. https://doi.org/10.1145/346988

Grigorev A, Sevastopolsky A, Vakhitov A, et al., 2019. Coordinate-based texture inpainting for pose-guided human image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12127-12136. https://doi.org/10.1109/CVPR.2019.01241

Gu Z, Li WB, Huo J, et al., 2021. LoFGAN: fusing local representations for few-shot image generation. Proc IEEE/CVF Int Conf on Computer Vision, p.8443-8451. https://doi.org/10.1109/ICCV48922.2021.00835

Guo XF, Yang HY, Huang D, 2021. Image inpainting via conditional texture and structure dual generation. Proc IEEE/CVF Int Conf on Computer Vision, p.14114-14123. https://doi.org/10.1109/ICCV48922.2021.01387

Habtegebrial TA, Jampani V, Gallo O, et al., 2020. Generative view synthesis: from single-view semantics to novel-view images. Proc 34th Int Conf on Neural Information Processing Systems, p.4745-4755.

Hall M, Bell SJ, Ross C, et al., 2024. Towards geographic inclusion in the evaluation of text-to-image models. ACM Conf on Fairness, Accountability, and Transparency, p.585-601. https://doi.org/10.1145/3630106.3658927

Hara T, Mukuta Y, Harada T, 2021. Spherical image generation from a single image by considering scene symmetry. *Proc AAAI Conf Artif Intell*, 35(2):1513-1521. https://doi.org/10.1609/aaai.v35i2.16242

Hassan MU, Alaliyat S, Hameed IA, 2023. Image generation models from scene graphs and layouts: a comparative analysis. *J King Saud Univ-Comput Inform Sci*, 35(5):101543. https://doi.org/10.1016/j.jksuci.2023.03.021

He S, Liao WT, Yang MY, et al., 2021. Context-aware layout to image generation with enhanced object appearance. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15044-15053. https://doi.org/10.1109/CVPR46437.2021.01480

He YT, Salakhutdinov R, Zico Kolter J, 2023. Localized text-to-image generation for free via cross attention control. https://arxiv.org/abs/2306.14636

Hessel J, Holtzman A, Forbes M, et al., 2021. CLIPScore: a reference-free evaluation metric for image captioning. Proc Conf on Empirical Methods in Natural Language Processing, p.7514-7528. https://doi.org/10.18653/v1/2021.emnlp-main.595

Heusel M, Ramsauer H, Unterthiner T, et al., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Proc 31st Int Conf on Neural Information Processing Systems, p.6626-6637.

Hinz T, Heinrich S, Wermter S, 2022. Semantic object accuracy for generative text-to-image synthesis. *IEEE Trans Patt Anal Mach Intell*, 44(3):1552-1565. https://doi.org/10.1109/TPAMI.2020.3021209

Ho J, Saharia C, Chan W, et al., 2022. Cascaded diffusion models for high fidelity image generation. *J Mach Learn Res*, 23(1):47.

Ho TT, Virtusio JJ, Chen YY, et al., 2020. Sketch-guided deep portrait generation. *ACM Trans Multim Comput Commun Appl*, 16(3):1-18. https://doi.org/10.1145/3396237

Hong S, Yan XC, Huang T, et al., 2018. Learning hierarchical semantic image manipulation through structured representations. Proc 32nd Int Conf on Neural Information Processing Systems, p.2713-2723.

Hong Y, Niu L, Zhang JF, et al., 2020. F2GAN: fusing-and-filling GAN for few-shot image generation. Proc 28th ACM Int Conf on Multimedia, p.2535-2543. https://doi.org/10.1145/3394171.3413561

Hu YS, Liu BL, Kasai J, et al., 2023. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. Proc IEEE/CVF Int Conf on Computer Vision, p.20349-20360. https://doi.org/10.1109/ICCV51070.2023.01866

Hua TY, Zheng HD, Bai YL, et al., 2021. Exploiting relationship for complex-scene image generation. *Proc AAAI Conf Artif Intell*, 35(2):1584-1592. https://doi.org/10.1609/aaai.v35i2.16250

Hua XS, Li J, 2015. Prajna: towards recognizing whatever you want from images without image labeling. *Proc AAAI Conf Artif Intell*, 29(1):137-144. https://doi.org/10.1609/aaai.v29i1.9186

Huang KY, Sun KY, Xie EZ, et al., 2023. T2I-compBench: a comprehensive benchmark for open-world compositional text-to-image generation. Proc 37th Int Conf on Neural Information Processing Systems, Article 3443.

Huang ST, Gong B, Feng YT, et al., 2024. Learning disentangled identifiers for action-customized text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7797-7806. https://doi.org/10.1109/CVPR52733.2024.00745

Huh M, Zhang R, Zhu JY, et al., 2020. Transforming and projecting images into class-conditional generative networks. 16th European Conf on Computer Vision, p.17-34. https://doi.org/10.1007/978-3-030-58536-5_2

Hulzebosch N, Ibrahimi S, Worring M, 2020. Detecting CNN-generated facial images in real-world scenarios. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.2729-2738. https://doi.org/10.1109/CVPRW50498.2020.00329

Ioannou E, Maddock S, 2024. Evaluation in neural style transfer: a review. *Comput Graph Forum*, 43(6):e15165. https://doi.org/10.1111/cgf.15165

Jayant NS, Noll P, 1984. Digital Coding of Waveforms: Principles and Applications to Speech and Video. Prentice-Hall, Englewood Cliffs, NJ, USA, p.139-140.

Jayasumana S, Ramalingam S, Veit A, et al., 2024. Rethinking FID: towards a better evaluation metric for image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9307-9315. https://doi.org/10.1109/CVPR52733.2024.00889

Jin D, Ma L, Liu RS, et al., 2021. Bridging the gap between low-light scenes: bilevel learning for fast adaptation. Proc 29th ACM Int Conf on Multimedia, p.2401-2409. https://doi.org/10.1145/3474085.3475404

Johnson J, Gupta A, Li FF, 2018. Image generation from scene graphs. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1219-1228. https://doi.org/10.1109/CVPR.2018.00133

Joo D, Kim D, Kim J, 2018. Generating a fusion image: one's identity and another's shape. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1635-1643. https://doi.org/10.1109/CVPR.2018.00176

Khashabi D, Stanovsky G, Bragg J, et al., 2022. GENIE: toward reproducible and standardized human evaluation for text generation. Proc Conf on Empirical Methods in Natural Language Processing, p.11444-11458. https://doi.org/10.18653/v1/2022.emnlp-main.787

Kim T, Song G, Lee S, et al., 2022. L-Verse: bidirectional generation between image and text. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16505-16515. https://doi.org/10.1109/CVPR52688.2022.01603

Kirstain Y, Polyak A, Singer U, et al., 2023. Pick-a-Pic: an open dataset of user preferences for text-to-image generation. Proc 37th Int Conf on Neural Information Processing Systems, p.36652-36663.

Klie JC, de Castilho RE, Gurevych I, 2024. Analyzing dataset annotation quality management in the wild. *Comput Ling*, 50(3):817-866. https://doi.org/10.1162/coli_a_00516

Koley S, Bhunia AK, Sain A, et al., 2023. Picture that sketch: photorealistic image generation from abstract sketches. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6850-6861. https://doi.org/10.1109/CVPR52729.2023.00662

Krizhevsky A, 2009. Learning Multiple Layers of Features from Tiny Images. MS Thesis, University of Toronto, Toronto, Canada.

Ku M, Li T, Zhang K, et al., 2024. ImagenHub: standardizing the evaluation of conditional image generation models. https://arxiv.org/abs/2310.01596

Kynkäänniemi T, Karras T, Laine S, et al., 2019. Improved precision and recall metric for assessing generative models. Proc 33rd Int Conf on Neural Information Processing Systems, p.3927-3936.

Kynkäänniemi T, Karras T, Aittala M, et al., 2023. The role of ImageNet classes in Fréchet inception distance. https://arxiv.org/abs/2203.06026

Ledig C, Theis L, Huszár F, et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.105-114. https://doi.org/10.1109/CVPR.2017.19

Lee T, Yasunaga M, Meng CL, et al., 2023. Holistic evaluation of text-to-image models. Proc 37th Int Conf on Neural Information Processing Systems, p.69981-70011.

Li BY, Ren WQ, Fu DP, et al., 2019. Benchmarking single-image dehazing and beyond. *IEEE Trans Image Process*, 28(1):492-505. https://doi.org/10.1109/TIP.2018.2867951

Li G, Zhao XF, Cao Y, et al., 2022. Manipulated face detection and localization based on semantic segmentation. Int Workshop on Digital Watermarking, p.98-113. https://doi.org/10.1007/978-3-031-25115-3_7

Li H, Shen CZ, Torr P, et al., 2024. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12006-12016. https://doi.org/10.1109/CVPR52733.2024.01141

Li HL, Pan SJ, Wang SQ, et al., 2018. Domain generalization with adversarial feature learning. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5400-5409. https://doi.org/10.1109/CVPR.2018.00566

Li J, Yu KW, Zhao YF, et al., 2019. Cross-reference stitching quality assessment for 360° omnidirectional images. Proc 27th ACM Int Conf on Multimedia, p.2360-2368. https://doi.org/10.1145/3343031.3350973

Li JY, Wang N, Zhang LF, et al., 2020. Recurrent feature reasoning for image inpainting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7757-7765. https://doi.org/10.1109/CVPR42600.2020.00778

Li LJ, Wang G, Li FF, 2007. OPTIMOL: automatic online picture collection via incremental model learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1-8. https://doi.org/10.1109/CVPR.2007.383048

Li LX, Zhang Y, Wang SH, 2023. The Euclidean space is evil: hyperbolic attribute editing for few-shot image generation. Proc IEEE/CVF Int Conf on Computer Vision, p.22657-22667. https://doi.org/10.1109/ICCV51070.2023.02076

Li SK, Fu JL, Liu KY, et al., 2024. CosmicMan: a text-to-image foundation model for humans. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6955-6965. https://doi.org/10.1109/CVPR52733.2024.00664

Li YJ, Zhang R, Lu JC, et al., 2020. Few-shot image generation with elastic weight consolidation. Proc 34th Int Conf on Neural Information Processing Systems, p.15885-15896.

Li ZJ, Wu JY, Koh I, et al., 2021. Image synthesis from layout with locality-aware mask adaption. Proc IEEE/CVF Int Conf on Computer Vision, p.13799-13808. https://doi.org/10.1109/ICCV48922.2021.01356

Liang XD, Zhang H, Lin L, et al., 2018. Generative semantic manipulation with mask-contrasting GAN. Proc 15th European Conf on Computer Vision, p.558-573. https://doi.org/10.1007/978-3-030-01261-8_34

Liang YW, He JF, Li G, et al., 2024. Rich human feedback for text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.19401-19411. https://doi.org/10.1109/CVPR52733.2024.01835

Lin TY, Maire M, Belongie S, et al., 2014. Microsoft COCO: common objects in context. 13th European Conf on Computer Vision, p.740-755. https://doi.org/10.1007/978-3-319-10602-1_48

Lin ZQ, Pathak D, Li BQ, et al., 2025. Evaluating text-to-visual generation with image-to-text generation. 18th European Conf on Computer Vision, p.366-384. https://doi.org/10.1007/978-3-031-72673-6_20

Liu JX, Liu Q, 2024. R3CD: scene graph to image generation with relation-aware compositional contrastive control diffusion. *Proc AAAI Conf Artif Intell*, 38(4):3657-3665. https://doi.org/10.1609/aaai.v38i4.28155

Liu RS, Ma L, Ma TY, et al., 2022. Learning with nested scene modeling and cooperative architecture search for low-light vision. *IEEE Trans Patt Anal Mach Intell*, 45(5):5953-5969. https://doi.org/10.1109/TPAMI.2022.3212995

Liu YF, Qin ZC, Wan T, et al., 2018. Auto-painter: cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks. *Neurocomputing*, 311:78-87. https://doi.org/10.1016/j.neucom.2018.05.045

Lu JW, Wang H, Shao TJ, et al., 2022. Pose guided image generation from misaligned sources via residual flow based correction. *Proc AAAI Conf Artif Intell*, 36(2):1863-1871. https://doi.org/10.1609/aaai.v36i2.20080

Luan FJ, Paris S, Shechtman E, et al., 2017. Deep photo style transfer. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6997-7005. https://doi.org/10.1109/CVPR.2017.740

Lucic M, Kurach K, Michalski M, et al., 2018. Are GANs created equal? A large-scale study. Proc 32nd Int Conf on Neural Information Processing Systems, p.698-707.

Luo A, Zhang ZT, Wu JJ, et al., 2020. End-to-end optimization of scene layout. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3753-3762. https://doi.org/10.1109/CVPR42600.2020.00381

Lv Z, Li X, Li X, et al., 2021. Learning semantic person image generation by region-adaptive normalization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10801-10810. https://doi.org/10.1109/CVPR46437.2021.01066

Ma KD, Fang YM, 2021. Image quality assessment in the modern age. Proc 29th ACM Int Conf on Multimedia, p.5664-5666. https://doi.org/10.1145/3474085.3478870

Ma LQ, Jia X, Sun QR, et al., 2017. Pose guided person image generation. Proc 31st Int Conf on Neural Information Processing Systems, p.406-416.

Machajdik J, Hanbury A, 2010. Affective image classification using features inspired by psychology and art theory. Proc 18th ACM Int Conf on Multimedia, p.83-92. https://doi.org/10.1145/1873951.1873965

Marrinan T, Papka ME, 2021. Real-time omnidirectional stereo rendering: generating 360° surround-view panoramic images for comfortable immersive viewing. *IEEE Trans Vis Comput Graph*, 27(5):2587-2596. https://doi.org/10.1109/TVCG.2021.3067780

Meng FQ, Shao WQ, Luo LX, et al., 2024. PhyBench: a physical commonsense benchmark for evaluating text-to-image models. https://arxiv.org/abs/2406.11802

Miyake R, Matsukawa T, Suzuki E, 2024. Image generation from hyper scene graph with multiple types of trinomial hyperedges. *SN Comput Sci*, 5(5):624. https://doi.org/10.1007/s42979-024-02791-8

Mondal AK, Tiwary P, Singla P, et al., 2023. Few-shot cross-domain image generation via inference-time latent-code learning. 11th Int Conf on Learning Representations.

Naeem MF, Oh SJ, Uh Y, et al., 2020. Reliable fidelity and diversity metrics for generative models. 37th Int Conf on Machine Learning, p.7176-7185.

Nazeri K, Ng E, Ebrahimi M, 2018. Image colorization using generative adversarial networks. 10th Int Conf on Articulated Motion and Deformable Objects, p.85-94. https://doi.org/10.1007/978-3-319-94544-6_9

Odena A, Olah C, Shlens J, 2017. Conditional image synthesis with auxiliary classifier GANs. Int Conf on Machine Learning, p.2642-2651.

Oh C, Cho W, Chae Y, et al., 2022. BIPS: bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. Proc 17th European Conf on Computer Vision, p.352-371. https://doi.org/10.1007/978-3-031-19787-1_20

Ojha U, Li YJ, Lu JW, et al., 2021. Few-shot image generation via cross-domain correspondence. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10738-10747. https://doi.org/10.1109/CVPR46437.2021.01060

Otani M, Togashi R, Sawai Y, et al., 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14277-14286. https://doi.org/10.1109/CVPR52729.2023.01372

Pang YX, Lin JX, Qin T, et al., 2022. Image-to-image translation: methods and applications. *IEEE Trans Multim*, 24:3859-3881. https://doi.org/10.1109/TMM.2021.3109419

Park T, Liu MY, Wang TC, et al., 2019. Semantic image synthesis with spatially-adaptive normalization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2332-2341. https://doi.org/10.1109/CVPR.2019.00244

Parmar G, Zhang R, Zhu JY, 2022. On aliased resizing and surprising subtleties in GAN evaluation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11410-11420. https://doi.org/10.1109/CVPR52688.2022.01112

Petsiuk V, Siemenn AE, Surbehera S, et al., 2022. Human evaluation of text-to-image models on a multi-task benchmark. https://arxiv.org/abs/2211.12112

Phaphuangwittayakul A, Guo Y, Ying FL, 2022. Fast adaptive meta-learning for few-shot image generation. *IEEE Trans Multim*, 24:2205-2217. https://doi.org/10.1109/TMM.2021.3077729

Phung Q, Ge SW, Huang JB, 2024. Grounded text-to-image synthesis with attention refocusing. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7932-7942. https://doi.org/10.1109/CVPR52733.2024.00758

Qiao TT, Zhang J, Xu DQ, et al., 2019a. Learn, imagine and create: text-to-image generation from prior knowledge. Proc 33rd Int Conf on Neural Information Processing Systems, p.885-895.

Qiao TT, Zhang J, Xu DQ, et al., 2019b. MirrorGAN: learning text-to-image generation by redescription. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1505-1514. https://doi.org/10.1109/CVPR.2019.00160

Quan FN, Lang B, 2024. Boundary-aware GAN for multiple overlapping objects in layout-to-image generation. *Multim Syst*, 30(2):88. https://doi.org/10.1007/s00530-024-01287-y

Quan WZ, Zhang RS, Zhang Y, et al., 2022. Image inpainting with local and global refinement. *IEEE Trans Image Process*, 31:2405-2420. https://doi.org/10.1109/TIP.2022.3152624

Ramesh A, Dhariwal P, Nichol A, et al., 2022. Hierarchical text-conditional image generation with CLIP latents. https://arxiv.org/abs/2204.06125

Ravuri S, Vinyals O, 2019. Classification accuracy score for conditional generative models. Proc 33rd Int Conf on Neural Information Processing Systems, p.12268-12279.

Regmi K, Borji A, 2018. Cross-view image synthesis using conditional GANs. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3501-3510. https://doi.org/10.1109/CVPR.2018.00369

Ren YR, Li G, Liu S, et al., 2020. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Trans Image Process*, 29:8622-8635. https://doi.org/10.1109/TIP.2020.3018224

Saharia C, Chan W, Saxena S, et al., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Proc 36th Int Conf on Neural Information Processing Systems, p.36479-36494.

Sajjadi MSM, Bachem O, Lucic M, et al., 2018. Assessing generative models via precision and recall. Proc 32nd Int Conf on Neural Information Processing Systems, p.5228-5237.

Salimans T, Goodfellow I, Zaremba W, et al., 2016. Improved techniques for training GANs. Proc 30th Int Conf on Neural Information Processing Systems, p.2234-2242.

Saseendran A, Skubch K, Keuper M, 2021. Multi-class multi-instance count conditioned adversarial image generation. Proc IEEE/CVF Int Conf on Computer Vision, p.6742-6751. https://doi.org/10.1109/ICCV48922.2021.00669

Sauer A, Schwarz K, Geiger A, 2022. StyleGAN-XL: scaling styleGAN to large diverse datasets. ACM SIGGRAPH, Article 49. https://doi.org/10.1145/3528233.3530738

Schroff F, Criminisi A, Zisserman A, 2011. Harvesting image databases from the web. *IEEE Trans Patt Anal Mach Intell*, 33(4):754-766. https://doi.org/10.1109/TPAMI.2010.133

Shen GB, Wang LZ, Lin JT, et al., 2024. SG-Adapter: enhancing text-to-image generation with scene graph guidance. https://arxiv.org/abs/2405.15321

Shen W, Liu RJ, 2017. Learning residual images for face attribute manipulation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1225-1233. https://doi.org/10.1109/CVPR.2017.135

Sheynin S, Polyak A, Singer U, et al., 2024. Emu Edit: precise image editing via recognition and generation tasks. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8871-8879. https://doi.org/10.1109/CVPR52733.2024.00847

Shi HY, Wang L, Zheng NN, et al., 2022. Loss functions for pose guided person image generation. *Patt Recogn*, 122:108351. https://doi.org/10.1016/j.patcog.2021.108351

Shibata K, Araki S, Maeda K, et al., 2014. High-quality panoramic image generation using multiple PAL images. *Electr Commun Jpn*, 97(6):58-66. https://doi.org/10.1002/ecj.11563

Shocher A, Gandelsman Y, Mosseri I, et al., 2020. Semantic pyramid for image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7455-7464. https://doi.org/10.1109/CVPR42600.2020.00748

Siarohin A, Sangineto E, Lathuilière S, et al., 2018. Deformable GANs for pose-based human image generation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3408-3416. https://doi.org/10.1109/CVPR.2018.00359

Sun W, Wu TF, 2021. Learning layout and style reconfigurable GANs for controllable image synthesis. *IEEE Trans Patt Anal Mach Intell*, 44(9):5070-5087. https://doi.org/10.1109/TPAMI.2021.3078577

Sushko V, Schönfeld E, Zhang D, et al., 2021. You only need adversarial supervision for semantic image synthesis. https://arxiv.org/abs/2012.04781

Sylvain T, Zhang PC, Bengio Y, et al., 2021. Object-centric image generation from layouts. *Proc AAAI Conf Artif Intell*, 35(3):2647-2655. https://doi.org/10.1609/aaai.v35i3.16368

Tan ZT, Chai ML, Chen DD, et al., 2021. Diverse semantic image synthesis via probability distribution modeling. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7958-7967. https://doi.org/10.1109/CVPR46437.2021.00787

Tang H, Xu D, Sebe N, et al., 2019. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2412-2421. https://doi.org/10.1109/CVPR.2019.00252

Tang H, Xu D, Yan Y, et al., 2020a. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7867-7876. https://doi.org/10.1109/CVPR42600.2020.00789

Tang H, Bai S, Zhang L, et al., 2020b. XingGAN for person image generation. 16th European Conf on Computer Vision, p.717-734. https://doi.org/10.1007/978-3-030-58595-2_43

Tang H, Shao L, Torr PHS, et al., 2023. Local and global GANs with semantic-aware upsampling for image generation. *IEEE Trans Patt Anal Mach Intell*, 45(1):768-784. https://doi.org/10.1109/TPAMI.2022.3155989

Tang JL, Yuan Y, Shao TJ, et al., 2021. Structure-aware person image generation with pose decomposition and semantic correlation. *Proc AAAI Conf Artif Intell*, 35(3):2656-2664. https://doi.org/10.1609/aaai.v35i3.16369

Tao X, Gao HY, Shen XY, et al., 2018. Scale-recurrent network for deep image deblurring. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.8174-8182. https://doi.org/10.1109/CVPR.2018.00853

Torralba A, Efros AA, 2011. Unbiased look at dataset bias. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1521-1528. https://doi.org/10.1109/CVPR.2011.5995347

Tripathi S, Sridhar SN, Sundaresan S, et al., 2019a. Compact scene graphs for layout composition and patch retrieval. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.676-683. https://doi.org/10.1109/CVPRW.2019.00094

Tripathi S, Bhiwandiwalla A, Bastidas A, et al., 2019b. Using scene graph context to improve image generation. https://arxiv.org/abs/1901.03762

Volokitin A, Konukoglu E, van Gool L, 2020. Decomposing image generation into layout prediction and conditional synthesis. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.1530-1538. https://doi.org/10.1109/CVPRW50498.2020.00194

Wang J, Chen ZW, Yuan CF, et al., 2023. Hierarchical curriculum learning for no-reference image quality assessment. *Int J Comput Vis*, 131(11):3074-3093. https://doi.org/10.1007/s11263-023-01851-5

Wang JR, Duan HY, Liu J, et al., 2023. AIGCIQA2023: a large-scale image quality assessment database for AI generated images: from the perspectives of quality, authenticity and correspondence. CAAI Int Conf on Artificial Intelligence, p.46-57. https://doi.org/10.1007/978-981-99-9119-8_5

Wang L, Chen W, Yang WJ, et al., 2020. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8:63514-63537. https://doi.org/10.1109/ACCESS.2020.2982224

Wang S, Saharia C, Montgomery C, et al., 2023. Imagen Editor and EditBench: advancing and evaluating text-guided image inpainting. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18359-18369. https://doi.org/10.1109/CVPR52729.2023.01761

Wang Y, Tao X, Qi XJ, et al., 2018. Image inpainting via generative multi-column convolutional neural networks. Proc 32nd Int Conf on Neural Information Processing Systems, p.331-340.

Wang Y, He YN, Li YZ, et al., 2024. InternVid: a large-scale video-text dataset for multimodal understanding and generation. https://arxiv.org/abs/2307.06942

Wang YH, Wang Q, Zhang DY, 2022. Few-shot generation by modeling stereoscopic priors. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.2370-2374. https://doi.org/10.1109/ICASSP43922.2022.9746576

Wang Z, Simoncelli EP, Bovik AC, 2003. Multiscale structural similarity for image quality assessment. 37th Asilomar Conf on Signals, Systems & Computers, p.1398-1402. https://doi.org/10.1109/ACSSC.2003.1292216

Wang Z, Bovik AC, Sheikh HR, et al., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*, 13(4):600-612. https://doi.org/10.1109/TIP.2003.819861

Wang ZH, Chen J, Hoi SCH, 2021. Deep learning for image super-resolution: a survey. *IEEE Trans Patt Anal Mach Intell*, 43(10):3365-3387. https://doi.org/10.1109/TPAMI.2020.2982166

Wang ZJ, Qi XQ, Yuan K, et al., 2022. Self-supervised correlation mining network for person image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7693-7702. https://doi.org/10.1109/CVPR52688.2022.00755

Wang ZM, Li YX, Huang DL, et al., 2023. DeformSg2im: scene graph based multi-instance image generation with a deformable geometric layout. *Neurocomputing*, 558:126684. https://doi.org/10.1016/j.neucom.2023.126684

Wu JY, Li ZJ, Zhang SY, et al., 2022. Amodal layout completion in complex outdoor scenes. 2nd CAAI Int Conf on Artificial Intelligence, p.30-41. https://doi.org/10.1007/978-3-031-20497-5_3

Wu JY, Gan WS, Chen ZF, et al., 2023. AI-generated content (AIGC): a survey. https://arxiv.org/abs/2304.06632

Wu SS, Tang H, Jing XY, et al., 2023. Cross-view panorama image synthesis. *IEEE Trans Multim*, 25:3546-3559. https://doi.org/10.1109/TMM.2022.3162474

Wu XS, Sun KQ, Zhu F, et al., 2023a. Human preference score: better aligning text-to-image models with human preference. Proc IEEE/CVF Int Conf on Computer Vision, p.2096-2105. https://doi.org/10.1109/ICCV51070.2023.00200

Wu XS, Hao YM, Sun KQ, et al., 2023b. Human preference score v2: a solid benchmark for evaluating human preferences of text-to-image synthesis. https://arxiv.org/abs/2306.09341

Wu YZ, Wang XT, Li Y, et al., 2021. Towards vivid and diverse image colorization with generative color prior. Proc IEEE/CVF Int Conf on Computer Vision, p.14357-14366. https://doi.org/10.1109/ICCV48922.2021.01411

Wu ZB, Deng HG, Wang Q, et al., 2023. SketchScene: scene sketch to image generation with diffusion models. IEEE Int Conf on Multimedia and Expo, p.2087-2092. https://doi.org/10.1109/ICME55011.2023.00357

Xia WH, Yang YJ, Xue JH, 2021a. Cali-sketch: stroke calibration and completion for high-quality face image generation from human-like sketches. *Neurocomputing*, 460:256-265. https://doi.org/10.1016/j.neucom.2021.07.029

Xia WH, Yang YJ, Xue JH, et al., 2021b. TediGAN: text-guided diverse face image generation and manipulation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2256-2265. https://doi.org/10.1109/CVPR46437.2021.00229

Xie Y, Fu YW, Tai Y, et al., 2022. Learning to memorize feature hallucination for one-shot image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9120-9129. https://doi.org/10.1109/CVPR52688.2022.00892

Xu HR, Chen SY, Zhang Y, 2023. Magical Brush: a symbol-based modern Chinese painting system for novices. Proc CHI Conf on Human Factors in Computing Systems, Article 131. https://doi.org/10.1145/3544548.3581429

Xu JZ, Liu X, Wu YC, et al., 2023. ImageReward: learning and evaluating human preferences for text-to-image generation. Proc 37th Int Conf on Neural Information Processing Systems, Article 700.

Xu QQ, Huang QM, Yao Y, 2012. Online crowdsourcing subjective image quality assessment. Proc 20th ACM Int Conf on Multimedia, p.359-368. https://doi.org/10.1145/2393347.2393400

Xu T, Zhang PC, Huang QY, et al., 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1316-1324. https://doi.org/10.1109/CVPR.2018.00143

Yan K, Ji L, Wu CF, et al., 2022. Trace controlled text to image generation. 17th European Conf on Computer Vision, p.59-75. https://doi.org/10.1007/978-3-031-20059-5_4

Yang H, Zhang RM, Guo XB, et al., 2020. Towards photo-realistic virtual try-on by adaptively generating ↔ preserving image content. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.7847-7856. https://doi.org/10.1109/CVPR42600.2020.00787

Yang L, Zhang ZL, Song Y, et al., 2023. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput Surv*, 56(4):105. https://doi.org/10.1145/3626235

Yang S, Jiang LM, Liu ZW, et al., 2022. Unsupervised image-to-image translation with generative prior. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18311-18320. https://doi.org/10.1109/CVPR52688.2022.01779

Yang Y, Lin YQ, Liu H, et al., 2024. Position: towards implicit prompt for text-to-image models. 41st Int Conf on Machine Learning, Article 2320.

Yao YZ, Hua XS, Shen FM, et al., 2016. A domain robust approach for image dataset construction. Proc 24th ACM Int Conf on Multimedia, p.212-216. https://doi.org/10.1145/2964284.2967213

Yu JH, Xu YZ, Koh JY, et al., 2022. Scaling autoregressive models for content-rich text-to-image generation. https://arxiv.org/abs/2206.10789

Yu YC, Zhan FN, Lu SJ, et al., 2021. WaveFill: a wavelet-based generation network for image inpainting. Proc IEEE/CVF Int Conf on Computer Vision, p.14094-14103. https://doi.org/10.1109/ICCV48922.2021.01385

Yuan XD, Tang ST, Li KJ, et al., 2024. CamFreeDiff: camera-free image to panorama generation with diffusion model. https://arxiv.org/abs/2407.07174

Zhai GT, Min XK, 2020. Perceptual image quality assessment: a survey. *Sci China Inform Sci*, 63(11):211301. https://doi.org/10.1007/s11432-019-2757-1

Zhang C, Wu QY, Gambardella CC, et al., 2024. Taming stable diffusion for text to 360° panorama image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6347-6357. https://doi.org/10.1109/CVPR52733.2024.00607

Zhang F, Tian SL, Huang ZQ, et al., 2024. Evaluation agent: efficient and promptable evaluation framework for visual generative models. https://arxiv.org/abs/2412.09645

Zhang H, Koh JY, Baldridge J, et al., 2021. Cross-modal contrastive learning for text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.833-842. https://doi.org/10.1109/CVPR46437.2021.00089

Zhang HG, Dai YC, Li HD, et al., 2019. Deep stacked hierarchical multi-patch network for image deblurring. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5971-5979. https://doi.org/10.1109/CVPR.2019.00613

Zhang JM, Ma CX, Yang KL, et al., 2022. Transfer beyond the field of view: dense panoramic semantic segmentation via unsupervised domain adaptation. *IEEE Trans Intell Transp Syst*, 23(7):9478-9491. https://doi.org/10.1109/TITS.2021.3123070

Zhang KH, Ren WQ, Luo WH, et al., 2022. Deep image deblurring: a survey. *Int J Comput Vis*, 130(9):2103-2130. https://doi.org/10.1007/s11263-022-01633-5

Zhang PZ, Yang LX, Xie XH, et al., 2022. Lightweight texture correlation network for pose guided person image generation. *IEEE Trans Circ Syst Video Technol*, 32(7):4584-4598. https://doi.org/10.1109/TCSVT.2021.3131738

Zhang R, Isola P, Efros AA, 2016. Colorful image colorization. 14th European Conf on Computer, p.649-666. https://doi.org/10.1007/978-3-319-46487-9_40

Zhang R, Isola P, Efros AA, et al., 2018. The unreasonable effectiveness of deep features as a perceptual metric. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.586-595. https://doi.org/10.1109/CVPR.2018.00068

Zhang SX, Wang BH, Wu JQ, et al., 2024. Learning multi-dimensional human preference for text-to-image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8018-8027. https://doi.org/10.1109/CVPR52733.2024.00766

Zhang XB, Zhai DH, Li TR, et al., 2023. Image inpainting based on deep learning: a review. *Inform Fus*, 90:74-94. https://doi.org/10.1016/j.inffus.2022.08.033

Zhang YK, Meng CY, Li ZJ, et al., 2023. Learning object consistency and interaction in image generation from scene graphs. Proc 32nd Int Joint Conf on Artificial Intelligence, p.1731-1739. https://doi.org/10.24963/ijcai.2023/192

Zhao B, Meng LL, Yin WD, et al., 2019. Image generation from layout. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8576-8585. https://doi.org/10.1109/CVPR.2019.00878

Zhao K, Yuan K, Sun M, et al., 2023. Quality-aware pretrained models for blind image quality assessment. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22302-22313. https://doi.org/10.1109/CVPR52729.2023.02136

Zhao SY, Zhang L, Shen Y, et al., 2021. RefineDNet: a weakly supervised refinement framework for single image dehazing. *IEEE Trans Image Process*, 30:3391-3404. https://doi.org/10.1109/TIP.2021.3060873

Zhao Y, Ren DY, Chen Y, et al., 2022. Cartoon image processing: a survey. *Int J Comput Vis*, 130(11):2733-2769. https://doi.org/10.1007/s11263-022-01645-1

Zhao YQ, Ding HH, Huang HJ, et al., 2022. A closer look at few-shot image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9130-9140. https://doi.org/10.1109/CVPR52688.2022.00893

Zhao YQ, Chandrasegaran K, Abdollahzadeh M, et al., 2023. AdAM: few-shot image generation via adaptation-aware kernel modulation. https://arxiv.org/abs/2307.01465

Zheng BY, Gu JJ, Li SJ, et al., 2024. LM4LV: a frozen large language model for low-level vision tasks. https://arxiv.org/abs/2405.15734

Zheng WD, Teng JY, Yang ZY, et al., 2024. CogView3: finer and faster text-to-image generation via relay diffusion. 18th European Conf on Computer Vision, p.1-22. https://doi.org/10.1109/10.1007/978-3-031-72980-5_1

Zhou MQ, Wang YX, Hou J, et al., 2024. SceneX: procedural controllable large-scale scene generation. https://arxiv.org/abs/2403.15698

Zhou S, Gordon ML, Krishna R, et al., 2019. HYPE: a benchmark for human eye perceptual evaluation of generative models. Proc 33rd Int Conf on Neural Information Processing Systems, p.3449-3461.

Zhou YF, Liu BC, Zhu YZ, et al., 2023. Shifted diffusion for text-to-image generation. Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10157-10166. https://doi.org/10.1109/CVPR52729.2023.00979

Zhou YF, Zhang RY, Zheng KZ, et al., 2024. Toffee: efficient million-scale dataset construction for subject-driven text-to-image generation. https://arxiv.org/abs/2406.09305

Zhu JY, Ma HM, Chen JS, et al., 2024. High-quality and diverse few-shot image generation via masked discrimination. *IEEE Trans Image Process*, 33:2950-2965. https://doi.org/10.1109/TIP.2024.3385295

Zhu WH, Zhai GT, Hu MH, et al., 2018. Arrow's impossibility theorem inspired subjective image quality assessment approach. *Signal Process*, 145:193-201. https://doi.org/10.1016/j.sigpro.2017.12.001

Zhu Z, Huang TT, Shi BG, et al., 2019. Progressive pose attention transfer for person image generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2342-2351. https://doi.org/10.1109/CVPR.2019.00245

Zhu Z, Xu ZL, You AS, et al., 2020. Semantically multi-modal image synthesis. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5466-5475. https://doi.org/10.1109/CVPR42600.2020.00551