Novelty Assessment Report

Paper: ImagenWorld: Stress-Testing Image Generation Models with Explainable Human Evaluation on Open-ended Real-World Tasks

PDF URL: https://openreview.net/pdf?id=bld9g6jFh9

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Advances in diffusion, autoregressive, and hybrid models have enabled high-quality image synthesis for tasks such as text-to-image, editing, and reference-guided composition. Yet, existing benchmarks remain limited, either focus on isolated tasks, cover only narrow domains, or provide opaque scores without explaining failure modes. We introduce \textbf{ImagenWorld}, a benchmark of 3.6K condition sets spanning six core tasks (generation and editing, with single or multiple references) and six topical domains (artworks, photorealistic images, information graphics, textual graphics, computer graphics, and screenshots). The benchmark is supported by 20K fine-grained human annotations and an explainable evaluation schema that tags localized object-level and segment-level errors, complementing automated VLM-based metrics. Our large-scale evaluation of 14 models yields several insights: (1) models typically struggle more in editing tasks than in generation tasks, especially in local edits. (2) models excel in artistic and photorealistic settings but struggle with symbolic and text-heavy domains such as screenshots and information graphics. (3) closed-source systems lead overall, while targeted data curation (e.g., Qwen-Image) narrows the gap in text-heavy cases. (4) modern VLM-based metrics achieve Kendall accuracies up to 0.79, approximating human ranking, but fall short of fine-grained, explainable error attribution. ImagenWorld provides both a rigorous benchmark and a diagnostic tool to advance robust image generation.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Benchmarking Image Generation and Editing Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Evaluation Frameworks and Benchmarks**
- **Model Architectures and Training Approaches**
- **Controllable Generation and Semantic Manipulation**
- **Application Domains and Specialized Tasks**
- **Survey and Taxonomic Studies**
- **Supporting Techniques and Infrastructure**

Complete Taxonomy Tree

- Benchmarking Image Generation and Editing Models Survey Taxonomy
- Evaluation Frameworks and Benchmarks
  - Multi-Task and Multi-Domain Benchmarks ★ (3 papers)
  - [0] ImagenWorld: Stress-Testing Image Generation Models with Explainable Human Evaluation on Open-ended Real-World Tasks (Anon et al., 2026) View paper
  - [18] MMIG-Bench: Towards Comprehensive and Explainable Evaluation of Multi-Modal Image Generation Models (Hua Hang, 2025) View paper
  - [22] Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing (Chen Zhihong, 2025) View paper
  - Task-Specific Evaluation Benchmarks (8 papers)
  - [19] Evaluating Generative AI Models for Image-Text Modification (Jayesh Soni, 2025) View paper
  - [21] Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms (Zitong Xu, 2025) View paper
  - [27] EditInspector: A Benchmark for Evaluation of Text-Guided Image Edits (Yosef, 2025) View paper
  - [29] Adiee: Automatic dataset creation and scorer for instruction-guided image editing evaluation (Wei Yi, 2025) View paper
  - [30] RefEdit: A Benchmark and Method for Improving Instruction-based Image Editing Model on Referring Expressions (Pathiraja, 2025) View paper
  - [33] Complex-Edit: CoT-Like Instruction Generation for Complexity-Controllable Image Editing Benchmark (Yang, 2025) View paper
  - [43] KRIS-Bench: Benchmarking Next-Level Intelligent Image Editing Models (Wu Yongliang, 2025) View paper
  - [49] From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing (Jingxuan Wei, 2024) View paper
  - Evaluation Metrics and Human Alignment (1 papers)
  - [46] Robust Watermarking Using Generative Priors Against Image Editing: From Benchmarking to Advances (LU Shilin, 2024) View paper
- Model Architectures and Training Approaches
  - Diffusion-Based Generation and Editing (4 papers)
  - [2] Sdedit: Guided image synthesis and editing with stochastic differential equations (Chenlin Meng, 2021) View paper
  - [4] FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space (Black Forest Labs, 2025) View paper
  - [6] More control for free! image synthesis with semantic diffusion guidance (Xihui Liu, 2023) View paper

  - [24] MMaDA-Parallel: Multimodal Large Diffusion Language Models for Thinking-Aware Editing and Generation (Ye Tian, 2025) View paper
  - GAN-Based Architectures (6 papers)
  - [5] Stargan v2: Diverse image synthesis for multiple domains (Yunjey Choi, 2020) View paper
  - [12] Conditional image synthesis with auxiliary classifier gans (Odena, 2017) View paper
  - [17] Stargan: Unified generative adversarial networks for multi-domain image-to-image translation (Choi, 2018) View paper
  - [25] A Comparative Study on Image Translation GAN Models to Improve Object Detection Accuracy in Low-Resource Domains (Yash Kumar Sahu, 2024) View paper
  - [39] Coupled generative adversarial networks (Liu Ming-yu, 2016) View paper
  - [40] StudioGAN: a taxonomy and benchmark of GANs for image synthesis (Minguk Kang, 2023) View paper
  - Autoregressive and Hybrid Models (2 papers)
  - [36] PlanGen: Towards Unified Layout Planning and Image Generation in Auto-Regressive Vision Language Models (He RunZe, 2025) View paper
  - [38] EditVerse: Unifying Image and Video Editing and Generation with In-Context Learning (Ju, 2025) View paper
- Controllable Generation and Semantic Manipulation
  - Latent Space Manipulation and Disentanglement (3 papers)
  - [8] Designing an encoder for stylegan image manipulation (Omer Tov, 2021) View paper
  - [23] StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation (Zongze Wu, 2020) View paper
  - [41] SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing (Yichun Shi, 2021) View paper
  - Text and Multimodal Conditioning (4 papers)
  - [3] Styleclip: Text-driven manipulation of stylegan imagery (Patashnik, 2021) View paper
  - [9] Sound-guided semantic image manipulation (Seung Hyun Lee, 2022) View paper
  - [13] Semantic image synthesis via adversarial learning (Hao Dong, 2017) View paper
  - [16] Manigan: Text-guided image manipulation (Bowen Li, 2020) View paper
  - Spatial and Compositional Control (5 papers)
  - [14] Image synthesis from reconfigurable layout and style (Wei Sun, 2019) View paper
  - [32] CIMGEN: Controlled Image Manipulation by Finetuning Pretrained Generative Models on Limited Data (Gudavalli, 2024) View paper
  - [44] MIGC++: Advanced Multi-Instance Generation Controller for Image Synthesis (Dewei Zhou, 2024) View paper
  - [45] EliGen: Entity-Level Controlled Image Generation with Regional Attention (Zhang Hong, 2025) View paper
  - [50] PRISM: Progressive Restoration for Scene Graph-based Image Manipulation (Pavel Jahoda, 2023) View paper
- Application Domains and Specialized Tasks
  - Data Augmentation and Domain Adaptation (4 papers)
  - [7] Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions (Panagiotis Alimisis, 2025) View paper
  - [28] A survey on Image Data Augmentation for Deep Learning (Connor Shorten, 2019) View paper
  - [42] GenMix: Effective Data Augmentation with Generative Diffusion Model Image Editing (Islam, 2024) View paper
  - [48] Applications of Cut, Paste, Learn Synthetic Image Generation, and Convolutional Neural Networks in Marine Animals Classification (Lanz Vincent T. Vencer, 2023) View paper
  - Specialized Visual Content Generation (1 papers)
  - [31] Factuality Matters: When Image Generation and Editing Meet Structured Visuals (Zhuo Le, 2025) View paper
  - Video Generation and Editing (1 papers)
  - [20] OmniV2V: Versatile Video Generation and Editing via Dynamic Content Manipulation (Liang Sen, 2025) View paper
  - Visuomotor Control and Robotics (2 papers)
  - [11] Generative Image as Action Models (Shridhar, 2024) View paper
  - [35] EC-Diffuser: Multi-Object Manipulation via Entity-Centric Behavior Generation (Qi, 2024) View paper
- Survey and Taxonomic Studies (6 papers)
  - [1] Multimodal image synthesis and editing: A survey and taxonomy (F Zhan, 2023) View paper
  - [10] Deep generative adversarial networks for image-to-image translation: A review (Aziz Alotaibi, 2020) View paper
  - [15] Exploring Generative AI: Models, Applications, and Challenges in Data Synthesis (S. Ramalakshmi, 2024) View paper
  - [26] Generative AI in Vision: A Survey on Models, Metrics and Applications (Singh, 2024) View paper
  - [34] Text-to-image cross-modal generation: A systematic review (Å»elaszczyk, 2024) View paper
  - [37] A state-of-the-art review on image synthesis with generative adversarial networks (Suhas Jangoan -, 2020) View paper
- Supporting Techniques and Infrastructure (1 papers)
  - [47] Deep Image Synthesis, Analysis and Indexing Using Integrated CNN Architectures (Muhammad Arslan, 2024) View paper

Narrative

Core task: Benchmarking image generation and editing models across diverse tasks and domains. The field has evolved into a rich ecosystem organized around six major branches. Evaluation Frameworks and Benchmarks establish standardized testbeds for assessing model capabilities, ranging from single-task metrics to comprehensive multi-domain suites like ImagenWorld[0] and MMIG Bench[18]. Model Architectures and Training Approaches encompass foundational techniques from early GANs such as StarGAN[17] and StarGAN v2[5] to modern diffusion-based systems. Controllable Generation and Semantic Manipulation focuses on methods that enable fine-grained steering of outputs through text, layout, or semantic guidance, exemplified by works like StyleCLIP[3] and SDEdit[2]. Application Domains and Specialized Tasks address domain-specific challenges in areas such as medical imaging, fashion, and video synthesis. Survey and Taxonomic Studies, including Multimodal Synthesis Survey[1] and Generative Vision Survey[26], provide meta-level perspectives on the landscape. Supporting Techniques and Infrastructure covers auxiliary components like data augmentation strategies reviewed in Diffusion Augmentation Review[7] and Image Augmentation Survey[28].

A particularly active tension exists between comprehensive multi-task benchmarks and specialized evaluation frameworks. While some efforts pursue breadth—testing models on numerous generation and editing operations simultaneously—others prioritize depth in specific modalities or interaction paradigms, as seen in OpenGPT Image[22] and EditInspector[27]. ImagenWorld[0] sits within the Multi-Task and Multi-Domain Benchmarks cluster, emphasizing holistic assessment across varied scenarios rather than narrow task optimization. This contrasts with more focused benchmarks like MMIG Bench[18], which targets multimodal instruction-guided generation, and specialized editing evaluations such as Complex Edit[33]. The central challenge remains balancing coverage against evaluation granularity: broad benchmarks risk superficial assessment, while narrow ones may miss emergent capabilities. ImagenWorld[0] addresses this by spanning multiple domains and task types, positioning itself as a unifying testbed that captures both generation fidelity and editing versatility across the spectrum of contemporary model capabilities.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. MMIG-Bench: Towards Comprehensive and Explainable Evaluation of Multi-Modal Image Generation Models

Authors: Hua Hang, Zeng, Ziyun, Hang Hua, Song Yizhi, et al. (19 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Recent multimodal image generators such as GPT-4o, Gemini 2.0 Flash, and Gemini 2.5 Pro excel at following complex instructions, editing images and maintaining concept consistency. However, they are still evaluated by disjoint toolkits: text-to-image (T2I) benchmarks that lacks multi-modal conditioning, and customized image generation benchmarks that overlook compositional semantics and common knowledge. We propose MMIG-Bench, a comprehensive Multi-Modal Image Generation Benchmark that unifies t...

Relationship Analysis

Both papers belong to the Multi-Task and Multi-Domain Benchmarks category, evaluating image generation and editing models across diverse tasks and visual domains. They overlap in their comprehensive evaluation approach, combining multiple generation/editing tasks with human evaluation and VLM-based metrics. However, ImagenWorld emphasizes explainable human evaluation with fine-grained error tagging (object-level and segment-level annotations) across 3.6K condition sets and 6 topical domains, while MMIG-Bench focuses on multi-modal conditioning with reference images (4,850 prompts paired with 1,750 multi-view reference images) and introduces the Aspect Matching Score (AMS) as a novel VQA-based metric for prompt-image alignment.

2. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing

Authors: Chen Zhihong, Bai Xue-hai, Zhihong Chen, Shi Yang, Xue-Yuan Bai, et al. (26 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The performance of unified multimodal models for image generation and editing is fundamentally constrained by the quality and comprehensiveness of their training data. While existing datasets have covered basic tasks like style transfer and simple object manipulation, they often lack the systematic structure and challenging scenarios required for real-world applications. To address this bottleneck, we introduce OpenGPT-4o-Image, a large-scale dataset constructed using a novel methodology that co...

Relationship Analysis

Both papers belong to the Multi-Task and Multi-Domain Benchmarks category, evaluating image generation and editing models across diverse tasks and domains. They overlap in covering multiple generation and editing tasks (text-to-image, reference-guided generation, instruction-based editing) with human evaluation and VLM-based metrics. However, ImagenWorld focuses on stress-testing with explainable human evaluation featuring fine-grained error attribution (object-level and segment-level tagging), while OpenGPT-4o-Image emphasizes dataset construction through automated pipelines and hierarchical taxonomy for training data creation rather than evaluation.

Contributions Analysis

Overall novelty summary. The paper introduces ImagenWorld, a benchmark spanning six core tasks (generation and editing with single or multiple references) across six topical domains, supported by 3.6K condition sets and 20K human annotations. Within the taxonomy, it resides in the 'Multi-Task and Multi-Domain Benchmarks' leaf alongside two sibling papers. This leaf is relatively sparse, containing only three papers total, suggesting that comprehensive multi-task, multi-domain evaluation frameworks remain an underexplored area compared to the broader field of 50 papers across 15 leaf nodes.

The taxonomy reveals that most evaluation work concentrates in adjacent leaves: 'Task-Specific Evaluation Benchmarks' contains eight papers focusing on narrow editing tasks, while 'Evaluation Metrics and Human Alignment' holds one paper on automated metrics. The sibling papers in ImagenWorld's leaf address multimodal instruction-guided generation and general multi-task assessment, but the scope notes clarify that ImagenWorld's simultaneous coverage of diverse visual domains (artworks, screenshots, information graphics) distinguishes it from single-domain or single-task approaches. Neighboring branches like 'Controllable Generation' and 'Model Architectures' focus on methods rather than evaluation infrastructure.

Among 30 candidates examined, none clearly refute any of the three contributions. Contribution A (diverse tasks and domains) examined 10 candidates with 0 refutable; Contribution B (large-scale human study revealing failure modes) examined 10 with 0 refutable; Contribution C (explainable evaluation schema with localized error attribution) examined 10 with 0 refutable. The statistics suggest that within this limited search scope, the combination of multi-task coverage, domain diversity, and fine-grained error tagging appears relatively novel, though the small candidate pool means substantial prior work may exist beyond the top-30 semantic matches.

Based on the limited literature search (30 candidates from semantic retrieval), the work appears to occupy a sparsely populated niche within evaluation frameworks. The taxonomy structure shows that while task-specific benchmarks are common, comprehensive multi-domain testbeds with explainable error attribution are less prevalent. However, the analysis does not cover exhaustive citation networks or domain-specific venues, so definitive claims about novelty require broader investigation beyond the top-K matches examined here.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: ImagenWorld benchmark with diverse tasks and domains

Description: The authors present ImagenWorld, a benchmark comprising 3.6K condition sets that systematically covers six representative task types (generation and editing with single or multiple references) and six topical domains (artworks, photorealistic images, information graphics, textual graphics, computer graphics, and screenshots), providing a unified testbed for evaluating image generation models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space

URL: View paper

Brief Assessment

FLUX Kontext[4] introduces KontextBench with 1026 pairs covering five task categories (local/global editing, character/style reference, text editing), which differs from ImagenWorld's systematic coverage of six task types and six topical domains with 3.6K condition sets and explainable human evaluation schema.

2. Ice-bench: A unified and comprehensive benchmark for image creating and editing

URL: View paper

Brief Assessment

Ice Bench[53] focuses on a different task decomposition (creating vs. editing with source/reference images) and emphasizes controllability metrics, whereas ImagenWorld emphasizes explainable human evaluation with object-level and segment-level error tagging across six topical domains (artworks, photorealistic, information graphics, textual graphics, computer graphics, screenshots).

3. UPGPT: Universal Diffusion Model for Person Image Generation, Editing and Pose Transfer

URL: View paper

Brief Assessment

UPGPT[59] focuses on person image generation, editing, and pose transfer tasks, not on creating a benchmark for evaluating image generation models across diverse domains and task types.

4. UniVG: A Generalist Diffusion Model for Unified Image Generation and Editing

URL: View paper

Brief Assessment

UniVG[55] focuses on building a unified generalist diffusion model for multiple image generation tasks, not on creating a benchmark for evaluating such models. The candidate does not present a benchmark dataset with human annotations or evaluation protocols.

5. OmniGen: Unified Image Generation

URL: View paper

Brief Assessment

OmniGen[51] presents a unified model architecture for image generation, not a benchmark. ImagenWorld's contribution is a comprehensive evaluation benchmark with 3.6K condition sets across six tasks and six domains, while OmniGen[51] focuses on model unification capabilities.

6. Anyedit: Mastering unified high-quality image editing for any idea

URL: View paper

Brief Assessment

AnyEdit[54] focuses on instruction-based image editing with 25 editing types across 5 domains, while ImagenWorld evaluates both generation and editing tasks across 6 task types and 6 topical domains with explainable human evaluation. The datasets serve different purposes and scopes.

7. DreamOmni: Unified Image Generation and Editing

URL: View paper

Brief Assessment

DreamOmni[52] focuses on building a unified model architecture for generation and editing tasks, not on creating a benchmark for evaluating such models. The candidate does not present a systematic evaluation framework across diverse domains.

8. UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics

URL: View paper

Brief Assessment

UniReal[57] focuses on a unified framework for image generation/editing by treating tasks as discontinuous video generation, not on creating a benchmark for evaluating models across diverse tasks and domains.

9. Mix mstar: A synthetic benchmark dataset for multi-class rotation vehicle detection in large-scale sar images

URL: View paper

Brief Assessment

Mix MSTAR[58] focuses exclusively on SAR vehicle detection in radar imagery, not on unified image generation and editing tasks across diverse visual domains like artworks, screenshots, and graphics.

10. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis

URL: View paper

Brief Assessment

Human Preference Score[56] focuses on evaluating human preferences for text-to-image synthesis quality through preference prediction, not on creating a unified benchmark spanning multiple generation and editing tasks across diverse domains.

Contribution 2: Large-scale human study revealing model failure modes

Description: The authors perform a comprehensive human evaluation study supported by 20K fine-grained annotations across 14 models, uncovering systematic failure patterns such as distinct editing biases, struggles with text-heavy domains, and performance gaps between closed-source and open-source systems.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Interactive Visual Assessment for Text-to-Image Generation Models

URL: View paper

Brief Assessment

Interactive Visual Assessment[67] focuses on dynamic interactive testing with LLM-powered adaptive prompt generation for text-to-image models, not a static large-scale human evaluation study with 20K annotations across 14 models revealing systematic failure patterns.

2. Benchmarking spatial relationships in text-to-image generation

URL: View paper

Brief Assessment

Spatial Relationships Benchmark[65] focuses on spatial relationship generation in text-to-image models, not general image generation failure modes. The human study evaluates spatial correctness rather than the broader failure patterns examined in the original paper.

3. Quality assessment for text-to-image generation: A survey

URL: View paper

Brief Assessment

Quality Assessment Survey[66] is a survey paper reviewing existing quality assessment methods for text-to-image generation. It does not present original human evaluation studies or reveal specific model failure modes through empirical investigation.

4. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment

URL: View paper

Brief Assessment

Linguistic Binding[68] focuses on attribute-entity binding failures in text-to-image generation through attention map alignment, not on comprehensive human evaluation studies revealing systematic failure patterns across multiple models and domains.

5. New Job, New Gender? Measuring the Social Bias in Image Generation Models

URL: View paper

Brief Assessment

Social Bias Generation[64] focuses on measuring social bias in image generation models through human evaluation of bias detection accuracy, not on revealing systematic failure patterns across diverse generation/editing tasks as in the original paper.

6. Rich human feedback for text-to-image generation

URL: View paper

Brief Assessment

Rich Human Feedback[60] focuses on collecting fine-grained human annotations (heatmaps, misaligned keywords) for text-to-image generation quality assessment, not on systematic evaluation across multiple models and tasks to reveal failure patterns as in the original paper.

7. Re-imagen: Retrieval-augmented text-to-image generator

URL: View paper

Brief Assessment

Re Imagen[63] focuses on retrieval-augmented text-to-image generation for rare entities, not on comprehensive human evaluation studies revealing systematic failure patterns across multiple models and tasks as in the original paper.

8. Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation

URL: View paper

Brief Assessment

Adversarial Nibbler[69] focuses on red-teaming for safety vulnerabilities in text-to-image models through adversarial prompts, not on comprehensive human evaluation revealing systematic failure patterns across diverse generation and editing tasks.

9. Imagenhub: Standardizing the evaluation of conditional image generation models

URL: View paper

Brief Assessment

ImagenHub[61] focuses on standardizing evaluation across seven conditional image generation tasks with human ratings for semantic consistency and perceptual quality. While it includes human evaluation (20K annotations across models), it does not systematically analyze failure modes with explainable annotations like object-level and segment-level error tagging as the original paper does.

10. Rethinking FID: Towards a Better Evaluation Metric for Image Generation

URL: View paper

Brief Assessment

Rethinking FID[62] focuses on evaluation metrics (FID vs CMMD) for image generation quality assessment, not on conducting human studies to reveal systematic failure patterns in image generation models. The candidate's human evaluation is used to validate metric reliability, not to uncover model failure modes.

Contribution 3: Explainable evaluation schema with localized error attribution

Description: The authors introduce a structured evaluation framework where human annotators not only provide scores but also tag specific failure modes with textual descriptions and localized masks at both object and segment levels, enabling interpretable diagnosis of model weaknesses beyond opaque numerical metrics.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models: A Case Study on ChatGPT

URL: View paper

Brief Assessment

Error Analysis Prompting[73] focuses on machine translation evaluation using LLM prompting strategies that emulate MQM error categorization, not on image generation evaluation with object-level and segment-level visual error masks.

2. PlanT: Explainable Planning Transformers via Object-Level Representations

URL: View paper

Brief Assessment

PlanT[70] focuses on explainability in autonomous driving planning through attention-based object prioritization, not on evaluation schemas for image generation models with human-annotated error attribution at object and segment levels.

3. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving

URL: View paper

Brief Assessment

Driving with LLMs[71] focuses on autonomous driving with QA evaluation metrics, not on image generation evaluation with localized object-level and segment-level error attribution.

4. Behaviour Discovery and Attribution for Explainable Reinforcement Learning

URL: View paper

Brief Assessment

Behaviour Discovery Attribution[74] focuses on explaining RL agent decisions through behavior segmentation and clustering, not on evaluating image generation models with localized object-level and segment-level error attribution as in the original paper.

5. Interpretable and accurate fine-grained recognition via region grouping

URL: View paper

Brief Assessment

Region Grouping Recognition[72] focuses on interpretable fine-grained recognition through part segmentation for classification tasks, not on evaluation methodologies with human-annotated error attribution for generative models.

6. Shap-based interpretable object detection method for satellite imagery

URL: View paper

Brief Assessment

SHAP Object Detection[76] focuses on feature attribution for object detection in satellite imagery to explain model inference, not on human evaluation schemas with localized error tagging for image generation models.

7. Revealing hidden context bias in segmentation and object detection through concept-specific explanations

URL: View paper

Brief Assessment

Hidden Context Bias[77] focuses on explaining segmentation and object detection models through concept-specific explanations to reveal biases, not on human evaluation frameworks for generative image models. The technical domains and objectives are fundamentally different.

8. Interpreting Object-level Foundation Models via Visual Precision Search

URL: View paper

Brief Assessment

Visual Precision Search[79] focuses on interpreting object-level foundation models (e.g., Grounding DINO, Florence-2) through attribution maps for visual grounding and object detection tasks. The original paper introduces an evaluation schema for image generation models with human-annotated object-level and segment-level error tags. These are fundamentally different domains and methodologies.

9. Odexai: A comprehensive object detection explainable ai evaluation

URL: View paper

Brief Assessment

OdexAI[75] focuses on evaluating XAI methods for object detection models using localization accuracy, faithfulness, and complexity metrics. It does not address human evaluation schemas for image generation models with object-level and segment-level error tagging as described in the original paper.

10. Explaining 3D Object Detection Through Shapley Value-Based Attribution Map

URL: View paper

Brief Assessment

Shapley 3D Detection[78] focuses on explaining 3D object detection models through Shapley value-based attribution maps for point clouds, not on creating an evaluation framework with human-annotated error tags for image generation models.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Imagenhub: Standardizing the evaluation of conditional image generation models

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] ImagenWorld: Stress-Testing Image Generation Models with Explainable Human Evaluation on Open-ended Real-World Tasks View paper
- [1] Multimodal image synthesis and editing: A survey and taxonomy View paper
- [2] Sdedit: Guided image synthesis and editing with stochastic differential equations View paper
- [3] Styleclip: Text-driven manipulation of stylegan imagery View paper
- [4] FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space View paper
- [5] Stargan v2: Diverse image synthesis for multiple domains View paper
- [6] More control for free! image synthesis with semantic diffusion guidance View paper
- [7] Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions View paper
- [8] Designing an encoder for stylegan image manipulation View paper
- [9] Sound-guided semantic image manipulation View paper
- [10] Deep generative adversarial networks for image-to-image translation: A review View paper
- [11] Generative Image as Action Models View paper
- [12] Conditional image synthesis with auxiliary classifier gans View paper

- [13] Semantic image synthesis via adversarial learning View paper
- [14] Image synthesis from reconfigurable layout and style View paper
- [15] Exploring Generative AI: Models, Applications, and Challenges in Data Synthesis View paper
- [16] Manigan: Text-guided image manipulation View paper
- [17] Stargan: Unified generative adversarial networks for multi-domain image-to-image translation View paper
- [18] MMIG-Bench: Towards Comprehensive and Explainable Evaluation of Multi-Modal Image Generation Models View paper
- [19] Evaluating Generative AI Models for Image-Text Modification View paper
- [20] OmniV2V: Versatile Video Generation and Editing via Dynamic Content Manipulation View paper
- [21] Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms View paper
- [22] Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing View paper
- [23] StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation View paper
- [24] MMaDA-Parallel: Multimodal Large Diffusion Language Models for Thinking-Aware Editing and Generation View paper
- [25] A Comparative Study on Image Translation GAN Models to Improve Object Detection Accuracy in Low-Resource Domains View paper
- [26] Generative AI in Vision: A Survey on Models, Metrics and Applications View paper
- [27] EditInspector: A Benchmark for Evaluation of Text-Guided Image Edits View paper
- [28] A survey on Image Data Augmentation for Deep Learning View paper
- [29] Adiee: Automatic dataset creation and scorer for instruction-guided image editing evaluation View paper
- [30] RefEdit: A Benchmark and Method for Improving Instruction-based Image Editing Model on Referring Expressions View paper
- [31] Factuality Matters: When Image Generation and Editing Meet Structured Visuals View paper
- [32] CIMGEN: Controlled Image Manipulation by Finetuning Pretrained Generative Models on Limited Data View paper
- [33] Complex-Edit: CoT-Like Instruction Generation for Complexity-Controllable Image Editing Benchmark View paper
- [34] Text-to-image cross-modal generation: A systematic review View paper
- [35] EC-Diffuser: Multi-Object Manipulation via Entity-Centric Behavior Generation View paper
- [36] PlanGen: Towards Unified Layout Planning and Image Generation in Auto-Regressive Vision Language Models View paper
- [37] A state-of-the-art review on image synthesis with generative adversarial networks View paper
- [38] EditVerse: Unifying Image and Video Editing and Generation with In-Context Learning View paper
- [39] Coupled generative adversarial networks View paper
- [40] StudioGAN: a taxonomy and benchmark of GANs for image synthesis View paper
- [41] SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing View paper
- [42] GenMix: Effective Data Augmentation with Generative Diffusion Model Image Editing View paper
- [43] KRIS-Bench: Benchmarking Next-Level Intelligent Image Editing Models View paper
- [44] MIGC++: Advanced Multi-Instance Generation Controller for Image Synthesis View paper
- [45] EliGen: Entity-Level Controlled Image Generation with Regional Attention View paper
- [46] Robust Watermarking Using Generative Priors Against Image Editing: From Benchmarking to Advances View paper
- [47] Deep Image Synthesis, Analysis and Indexing Using Integrated CNN Architectures View paper
- [48] Applications of Cut, Paste, Learn Synthetic Image Generation, and Convolutional Neural Networks in Marine Animals Classification View paper
- [49] From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing View paper
- [50] PRISM: Progressive Restoration for Scene Graph-based Image Manipulation View paper
- [51] OmniGen: Unified Image Generation View paper
- [52] DreamOmni: Unified Image Generation and Editing View paper
- [53] Ice-bench: A unified and comprehensive benchmark for image creating and editing View paper
- [54] Anyedit: Mastering unified high-quality image editing for any idea View paper
- [55] UniVG: A Generalist Diffusion Model for Unified Image Generation and Editing View paper
- [56] Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis View paper
- [57] UniReal: Universal Image Generation and Editing via Learning Real-world Dynamics View paper
- [58] Mix mstar: A synthetic benchmark dataset for multi-class rotation vehicle detection in large-scale sar images View paper
- [59] UPGPT: Universal Diffusion Model for Person Image Generation, Editing and Pose Transfer View paper
- [60] Rich human feedback for text-to-image generation View paper
- [61] Imagenhub: Standardizing the evaluation of conditional image generation models View paper
- [62] Rethinking FID: Towards a Better Evaluation Metric for Image Generation View paper
- [63] Re-imagen: Retrieval-augmented text-to-image generator View paper
- [64] New Job, New Gender? Measuring the Social Bias in Image Generation Models View paper
- [65] Benchmarking spatial relationships in text-to-image generation View paper
- [66] Quality assessment for text-to-image generation: A survey View paper
- [67] Interactive Visual Assessment for Text-to-Image Generation Models View paper
- [68] Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment View paper
- [69] Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation View paper
- [70] PlanT: Explainable Planning Transformers via Object-Level Representations View paper
- [71] Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving View paper
- [72] Interpretable and accurate fine-grained recognition via region grouping View paper
- [73] Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models: A Case Study on ChatGPT View paper
- [74] Behaviour Discovery and Attribution for Explainable Reinforcement Learning View paper
- [75] Odexai: A comprehensive object detection explainable ai evaluation View paper
- [76] Shap-based interpretable object detection method for satellite imagery View paper
- [77] Revealing hidden context bias in segmentation and object detection through concept-specific explanations View paper
- [78] Explaining 3D Object Detection Through Shapley Value-Based Attribution Map View paper
- [79] Interpreting Object-level Foundation Models via Visual Precision Search View paper