Novelty Assessment Report

Paper: Factuality Matters: When Image Generation and Editing Meet Structured Visuals
PDF URL: https://openreview.net/pdf?id=J1Rorvw7DQ
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

While modern visual generation models excel at creating aesthetically pleasing natural images, they struggle with producing or editing structured visuals like charts, diagrams, and mathematical figures, which demand composition planning, text rendering, and multimodal reasoning for factual fidelity. To address this, we present the first comprehensive, systematic investigation of this domain, encompassing data construction, model training, and an evaluation benchmark. First, we construct a large-scale dataset of 1.3 million high-quality structured image pairs derived from executable drawing programs and augmented with chain-of-thought reasoning annotations. Leveraging this dataset, we train a unified model that integrates a multimodal language model with FLUX.1-Kontext via a lightweight connector for enhanced multimodal understanding. A three-stage training curriculum enables progressive feature alignment, knowledge infusion, and reasoning-augmented generation, further boosted by an external reasoner at inference time. Finally, we introduce StructBench, a novel benchmark for generation and editing with over 2,000 challenging samples, and an accompanying evaluation metric, StructScore, which employs a multi-round Q&A protocol to assess fine-grained factual accuracy. Evaluations of 15 models reveal that even state-of-the-art systems score below 50\%, while our model achieves the strongest open-source performance, with consistent gains from inference-time reasoning. By releasing dataset, model, and benchmark, we aim to advance unified multimodal foundations for structured visuals.

Core Task Landscape

This paper addresses: **structured image generation and editing**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

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

- **Text-Guided Generation and Editing Frameworks**
- **Spatially-Constrained and Compositional Generation**
- **Subject-Driven and Consistency-Preserving Editing**
- **Structured Visual Generation with Factual Constraints**
- **Domain-Specific Synthesis with Structural Constraints**
- **Constrained Generation with Physical and Semantic Priors**
- **Auxiliary Capabilities and Supporting Technologies**

Complete Taxonomy Tree

- structured image generation and editing Survey Taxonomy
- Text-Guided Generation and Editing Frameworks
  - Open-Domain Text-to-Image Generation (4 papers)
  - [1] Vqgan-clip: Open domain image generation and editing with natural language guidance (Crowson, 2022) View paper
  - [15] More control for free! image synthesis with semantic diffusion guidance (Xihui Liu, 2023) View paper
  - [22] High fidelity text to image generation with contrastive alignment and structural guidance (Danyi Gao, 2025) View paper
  - [33] From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning (Zhuo Le, 2025) View paper
  - Unified Generation and Editing Models (2 papers)
  - [6] DreamOmni: Unified Image Generation and Editing (Xia Bin, 2025) View paper
  - [26] Controllable data generation by deep learning: A review (Shiyu Wang, 2024) View paper
  - Prompt-Based Image Editing (4 papers)
  - [24] Expressive Image Generation and Editing with Rich Text (Songwei Ge, 2025) View paper
  - [37] Prompt-to-Prompt Image Editing with Cross Attention Control (Hertz, 2022) View paper
  - [39] Describe, Don't Dictate: Semantic Image Editing with Natural Language Intent (Ci En, 2025) View paper
  - [40] Sequential attention GAN for interactive image editing (Yu Cheng, 2020) View paper
  - Interactive and Multi-Turn Editing (1 papers)
  - [31] CREA: A Collaborative Multi-Agent Framework for Creative Image Editing and Generation (Yanardag, 2025) View paper
- Spatially-Constrained and Compositional Generation
  - Layout-Guided and Box-Constrained Synthesis (4 papers)
  - [4] Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion (Jinheng Xie, 2023) View paper
  - [12] Composer: Creative and Controllable Image Synthesis with Composable Conditions (Huang Lianghua, 2023) View paper
  - [20] High-resolution image synthesis and semantic manipulation with conditional gans (Ting-Chun Wang, 2018) View paper
  - [42] Multi-region text-driven manipulation of diffusion imagery (Peng Zhou, 2024) View paper
  - Compositional Decomposition and Multi-Instance Control (4 papers)

    - [29] Training-free structured diffusion guidance for compositional text-to-image synthesis (Feng, 2022) View paper
    - [38] Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing (Yichun Shi, 2022) View paper
    - [45] Learning hierarchical semantic image manipulation through structured representations (Seunghoon Hong, 2018) View paper
    - [50] MIGC++: Advanced Multi-Instance Generation Controller for Image Synthesis (Dewei Zhou, 2024) View paper
    - Hierarchical and Modality-Cascaded Generation (3 papers)
    - [17] LayerCraft: Enhancing Text-to-Image Generation with CoT Reasoning and Layered Object Integration (ZHANG Yuyao, 2025) View paper
    - [27] ToddlerDiffusion: Interactive Structured Image Generation with Cascaded Schr\" odinger Bridge (Eslam Abdelrahman, 2023) View paper
    - [46] Image-POSER: Reflective RL for Multi-Expert Image Generation and Editing (Hossein Mohebbi, 2025) View paper
- Subject-Driven and Consistency-Preserving Editing
    - Subject Representation and Multimodal Control (3 papers)
    - [14] Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing (Li, 2023) View paper
    - [21] RAVEL: Rare Concept Generation and Editing via Graph-driven Relational Guidance (Kavana Venkatesh, 2024) View paper
    - [35] Dreamedit: Subject-driven image editing (Li Tianle, 2023) View paper
    - Attention-Based Consistency and Self-Attention Control (2 papers)
    - [13] MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing (Mingdeng Cao, 2023) View paper
    - [19] DrawingInStyles: Portrait image generation and editing with spatially conditioned StyleGAN (Su, 2022) View paper
    - Inversion-Based and Reconstruction-Preserving Editing (4 papers)
    - [7] Sdedit: Guided image synthesis and editing with stochastic differential equations (Chenlin Meng, 2021) View paper
    - [25] Generative visual manipulation on the natural image manifold (Zhu, 2016) View paper
    - [30] Pix2video: Video editing using image diffusion (Duygu Ceylan, 2023) View paper
    - [34] FireFlow: Fast Inversion of Rectified Flow for Image Semantic Editing (Deng Yingying, 2024) View paper
- Structured Visual Generation with Factual Constraints
    - Diagram and Chart Generation from Text (1 papers)
    - [11] From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing (Jingxuan Wei, 2024) View paper
    - Factual Fidelity and Reasoning-Augmented Generation ★ (1 papers)
    - [0] Factuality Matters: When Image Generation and Editing Meet Structured Visuals (Anon et al., 2026) View paper
    - Scene Graph-Based and Relational Guidance (1 papers)
    - [8] A Comprehensive Survey of Scene Graphs: Generation and Application (Xiaojun Chang, 2021) View paper
- Domain-Specific Synthesis with Structural Constraints
    - Medical Image Synthesis and Cross-Modality Translation (5 papers)
    - [9] Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN (Heran Yang, 2020) View paper
    - [16] PST-Diff: achieving high-consistency stain transfer by diffusion models with pathological and structural constraints (Yufang He, 2024) View paper
    - [18] Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN (Yang He-ran, 2018) View paper
    - [23] CT to ultrasound abdominal image synthesis using domain-constraint generative adversarial networks (Xiaojie Wang, 2025) View paper
    - [47] MED-INPAINT: Medical image synthesis using multi-level conditional inpainting with a denoising diffusion probabilistic model and adaptive contrast priors (Reza Kalantar, 2023) View paper
    - Industrial and Specialized Domain Synthesis (3 papers)
    - [5] Unpaired Synthesis of IC Scanning Electron Microscopy Images with Structural Constraints (LuYun Li, 2024) View paper
    - [36] Shape-controlled synthesis of silver and gold nanostructures (Benjamin Wiley, 2005) View paper
    - [49] A Novel Approach to Industrial Defect Generation through Blended Latent Diffusion Model with Online Adaptation (Li Hanxi, 2024) View paper
    - 3D and Multi-View Generation (2 papers)
    - [2] Structured 3D Latents for Scalable and Versatile 3D Generation (Jianfeng Xiang, 2024) View paper
    - [3] SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion (Voleti, 2024) View paper
- Constrained Generation with Physical and Semantic Priors
    - Structure-Preserving and Semantic Consistency (2 papers)
    - [41] Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint (Jiaxian Guo, 2022) View paper
    - [43] Image inpainting based on structural constraint and multi-scale feature fusion (Yao Fan, 2023) View paper
    - Physical and Optimization-Based Constraints (1 papers)
    - [48] Constrained synthesis with projected diffusion models (Stephen Baek, 2024) View paper
    - Adaptive Priors and Contrast-Guided Synthesis (1 papers)
    - [44] WeatherEdit: Controllable Weather Editing with 4D Gaussian Field (Qian, 2025) View paper
- Auxiliary Capabilities and Supporting Technologies
    - Provenance and Watermarking (1 papers)
    - [32] GenPTW: In-Generation Image Watermarking for Provenance Tracing and Tamper Localization (Liu Chun-ya, 2025) View paper
    - Controllability Surveys and Taxonomies (1 papers)
    - [10] Controllable image synthesis methods, applications and challenges: a comprehensive survey (Shanshan Huang, 2024) View paper
    - Image Indexing and Retrieval (1 papers)
    - [28] Deep Image Synthesis, Analysis and Indexing Using Integrated CNN Architectures (Muhammad Arslan, 2024) View paper

Narrative

Core task: structured image generation and editing. The field organizes itself around several complementary branches that reflect different ways of imposing structure on generative models. Text-Guided Generation and Editing Frameworks (e.g., VQGAN CLIP[1], Prompt to Prompt[37]) focus on leveraging natural language to steer synthesis, while Spatially-Constrained and Compositional

Generation (e.g., BoxDiff[4], Composer[12]) emphasizes explicit layout or bounding-box controls to arrange multiple objects. Subject-Driven and Consistency-Preserving Editing (e.g., MasaCtrl[13], BLIP Diffusion[14]) targets identity preservation and fine-grained attribute manipulation. Domain-Specific Synthesis with Structural Constraints addresses specialized applications such as medical imaging (e.g., CT Ultrasound Synthesis[23]) or industrial defect generation (e.g., Industrial Defect Generation[49]), where domain priors guide the output. Constrained Generation with Physical and Semantic Priors incorporates scene graphs, depth maps, or other intermediate representations (e.g., Scene Graphs Survey[8], Semantic Diffusion Guidance[15]) to enforce realism. Finally, Auxiliary Capabilities and Supporting Technologies provide foundational tools—such as 3D-aware latents (Structured 3D Latents[2], SV3D[3])—that enable richer structural control across these branches.

A particularly active line of work explores how to balance flexibility with fidelity: some methods prioritize training-free or plug-and-play guidance (e.g., Training Free Structured[29], Semantic Diffusion Guidance[15]), while others learn specialized modules for compositional reasoning (e.g., MIGC Plus[50], Composer[12]). Trade-offs between user control granularity and model complexity remain central, as do questions about how to verify that generated content respects factual or physical constraints. Factuality Structured Visuals[0] sits within the Structured Visual Generation with Factual Constraints branch, specifically under Factual Fidelity and Reasoning-Augmented Generation. It shares thematic ground with works like Text to Diagram[11], which also emphasizes correctness and structured reasoning, but distinguishes itself by foregrounding explicit factual verification mechanisms. Compared to domain-specific approaches such as IC SEM Synthesis[5], Factuality Structured Visuals[0] appears to target broader applicability across diverse factual scenarios, positioning it as a bridge between general-purpose text-guided frameworks and specialized constraint-driven synthesis.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on ensuring factual accuracy in structured visual outputs through explicit reasoning mechanisms (chain-of-thought, external reasoners, multimodal reasoning). The sibling subtopics address complementary aspects of structured generation: one targets diagram/chart creation from text with logical organization, while the other uses graph-based relational constraints to guide generation. All three share the goal of producing structured visual content, but differ in their primary mechanisms and quality assurances.

Similarities: - All three subtopics operate within structured image generation rather than open-domain natural image synthesis - Each emphasizes logical organization or relational coherence in visual outputs - All involve text-to-image or text-guided generation pipelines with structural constraints - Methods across subtopics aim to produce editable or interpretable visual representations

Differences: - The original leaf uniquely requires explicit reasoning components (chain-of-thought, external reasoners) and factual verification mechanisms, while siblings may achieve structure through other means - Diagram/Chart Generation focuses on domain-specific visual formats (flowcharts, diagrams) with editability, whereas the original leaf applies reasoning to broader structured outputs - Scene Graph-Based methods use explicit graph representations (nodes/edges for objects/relationships) as the primary structural constraint, while the original leaf uses reasoning processes to ensure factual fidelity - The original leaf emphasizes verification and accuracy guarantees, while siblings prioritize structural organization and relational correctness without necessarily validating factual claims

Suggested Search Directions: - Investigate hybrid approaches combining scene graphs with reasoning-augmented verification for factually accurate relational generation - Explore how chain-of-thought reasoning could enhance diagram generation to ensure logical consistency in flowcharts and technical visualizations - Examine methods that bridge factual verification with graph-based constraints for knowledge-grounded structured image synthesis

Sibling Subtopics

- **Diagram and Chart Generation from Text** (leaves: 1, papers: 1)
- Scope: Methods producing structured diagrams, flowcharts, or visualizations directly from textual descriptions with logical organization and editability.
- Exclude: Excludes natural image synthesis or methods without diagram-specific constraints; see Open-Domain Generation or Scene Graph-Based Generation.
- **Scene Graph-Based and Relational Guidance** (leaves: 1, papers: 1)
- Scope: Techniques using scene graphs or knowledge graphs to represent object relationships and guide generation with structured relational constraints.
- Exclude: Excludes methods without explicit graph representations; see Layout-Guided Synthesis or Factual Fidelity.

Contributions Analysis

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

Contribution 1: Large-scale structured image dataset with chain-of-thought annotations

Description: The authors build a dataset of 1.3 million structured image pairs (charts, diagrams, mathematical figures) generated from executable code. Each sample includes both text-to-image prompts and editing instructions, along with GPT-5-generated chain-of-thought reasoning trajectories that provide explicit analysis and planning steps.

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. AnnotatedTables: A Large Tabular Dataset with Language Model Annotations

URL: View paper

Brief Assessment

AnnotatedTables[62] focuses on tabular data annotation with SQL programs and input-target columns, not structured images (charts, diagrams, mathematical figures) with chain-of-thought reasoning for visual generation and editing tasks.

---

2. Benchmarking multimodal cot reward model stepwise by visual program

URL: View paper

Brief Assessment

Multimodal CoT Reward[59] focuses on training reward models for evaluating multimodal chain-of-thought reasoning steps, not on constructing structured image datasets from executable code. Their dataset (svip-train) contains 7,948 samples with 20,000 steps derived from visual programming for reward model training, which differs fundamentally from the original paper's 1.3 million structured image pairs generated from drawing programs for image generation/editing tasks.

---

3. Chartgemma: Visual instruction-tuning for chart reasoning in the wild

URL: View paper

Brief Assessment

ChartGemma[57] focuses on chart-specific instruction-tuning data generated from chart images for visual reasoning tasks, not on structured images like diagrams and mathematical figures from executable code with GPT-generated chain-of-thought trajectories as described in the original paper.

---

4. MINT-CoT: Enabling Interleaved Visual Tokens in Mathematical Chain-of-Thought Reasoning

URL: View paper

Brief Assessment

MINT CoT[64] focuses on mathematical reasoning with visual token selection at the token level, not on structured image generation/editing from executable code. The datasets serve fundamentally different purposes: MINT CoT[64] annotates mathematical problems with visual token indices for reasoning, while the original paper constructs image pairs from drawing programs for generation/editing tasks.

---

5. Simple o3: Towards Interleaved Vision-Language Reasoning

URL: View paper

Brief Assessment

Simple O3[61] focuses on interleaved vision-language reasoning with tool interactions (cropping, zooming) for multimodal tasks, not on structured image generation from executable code. Their TWI-Tools-146K dataset targets dynamic visual operations during reasoning chains, whereas the original paper constructs 1.3M structured images (charts, diagrams, math figures) from executable drawing programs with GPT-5 CoT annotations for generation and editing tasks.

---

6. Embodiedgpt: Vision-language pre-training via embodied chain of thought

URL: View paper

Brief Assessment

EmbodiedGPT[56] focuses on egocentric video datasets for embodied AI tasks (robot manipulation, planning) rather than structured visual images like charts, diagrams, and mathematical figures. The datasets serve fundamentally different domains and purposes.

---

7. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images

URL: View paper

Brief Assessment

CodePlot CoT[60] focuses on mathematical visual reasoning problems requiring code-driven image generation during the solving process, not on structured image generation/editing tasks like charts and diagrams as in the original paper.

---

8. Star: A benchmark for situated reasoning in real-world videos

URL: View paper

Brief Assessment

STAR[55] focuses on video-based situated reasoning with action-centric hypergraphs for real-world human activities, not structured image generation from executable code with chain-of-thought annotations.

---

9. Measuring and improving chain-of-thought reasoning in vision-language models

URL: View paper

Brief Assessment

Chain of Thought[58] focuses on vision-language models' reasoning consistency through question-answering chains for visual inference tasks, not on generating structured images (charts, diagrams, math figures) from executable code with editing instructions.

---

10. Translating a visual lego manual to a machine-executable plan

URL: View paper

Brief Assessment

Visual LEGO Manual[63] focuses on translating LEGO assembly manuals into machine-executable plans through sequential pose estimation, not on creating structured image datasets with chain-of-thought annotations from executable programs.

Contribution 2: Unified model with three-stage progressive training curriculum

Description: The authors propose a three-stage training pipeline that progressively aligns multimodal features from Qwen-VL with FLUX.1-Kontext via a lightweight MLP connector, infuses structured-visual knowledge, and incorporates chain-of-thought reasoning to enable inference-time scaling with an external reasoner.

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. Aguvis: Unified pure vision agents for autonomous gui interaction

URL: View paper

Brief Assessment

Aguvis[66] uses a two-stage training pipeline (grounding, then planning & reasoning), not three stages. The original paper's three-stage curriculum (alignment, knowledge infusion, reasoning-augmented generation) for multimodal feature alignment with FLUX.1-Kontext differs fundamentally from Aguvis's approach of training GUI agents with QWen2-VL.

---

2. Mutex: Learning unified policies from multimodal task specifications

URL: View paper

Brief Assessment

Mutex[70] focuses on multimodal task specifications for robot manipulation (text, speech, video, images), not on structured visual generation. The three-stage training in Mutex involves masked modeling and cross-modal matching for robot policy learning, which is fundamentally different from the original paper's progressive alignment of multimodal features for structured image generation and editing with chain-of-thought reasoning.

---

3. TinyRS-R1: Compact Multimodal Language Model for Remote Sensing

URL: View paper

Brief Assessment

TinyRS-R1[72] uses a four-stage training pipeline (pre-training, instruction tuning, CoT fine-tuning, GRPO alignment) for remote sensing tasks, while the original paper proposes a three-stage curriculum (feature alignment, knowledge infusion, reasoning-augmented generation) for structured visual generation. The domain, task objectives, and architectural integration differ fundamentally.

---

4. Enhancing Spatial Reasoning in Multimodal Large Language Models through Reasoning-based Segmentation

URL: View paper

Brief Assessment

Spatial Reasoning Segmentation[69] focuses on 3D point cloud segmentation with a two-stage reasoning framework (reasoning prior learning and prior-guided refinement), not multimodal image generation with progressive feature alignment and reasoning-augmented generation as in the original paper.

---

5. Universal Visuo-Tactile Video Understanding for Embodied Interaction

URL: View paper

Brief Assessment

Visuo Tactile Understanding[73] focuses on visuo-tactile video understanding for embodied interaction, not structured visual generation. The three-stage training addresses different modalities (tactile-visual-language) and objectives (tactile perception) compared to the original paper's structured image generation tasks.

---

6. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning

URL: View paper

Brief Assessment

Unified Multimodal Reward[71] focuses on reward model training for evaluating vision outputs through chain-of-thought reasoning, not on multimodal feature alignment for image generation/editing as in the original paper.

---

7. Reasonrec: A reasoning-augmented multimodal agent for unified recommendation

URL: View paper

Brief Assessment

ReasonRec[68] focuses on recommendation systems with a three-stage training pipeline (unified alignment, hybrid visual learning, thinking enhancement) for multimodal recommendation tasks, not structured visual generation. The technical domains and objectives differ fundamentally.

---

8. Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning

URL: View paper

Brief Assessment

Lingshu[67] employs a four-stage training pipeline (medical shallow alignment, medical deep alignment, medical instruction tuning, and medical-oriented RL) for medical multimodal models, while the original paper focuses on structured visual generation with a three-stage curriculum (feature alignment, knowledge infusion, reasoning-augmented generation). The domains and objectives differ fundamentally.

---

9. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing

URL: View paper

Brief Assessment

Visual Drawing Reasoning[74] focuses on spatial reasoning through drawing operations (bounding boxes, auxiliary lines) with a three-stage framework (cold-start, reflective rejection sampling, RL). The original paper addresses structured visual generation (charts, diagrams) with multimodal alignment, knowledge infusion, and reasoning-augmented generation. These are fundamentally different tasks and training objectives.

---

10. Multimodal chain-of-thought reasoning in language models

URL: View paper

Brief Assessment

Multimodal Chain of Thought[65] proposes a two-stage framework (rationale generation and answer inference) for multimodal reasoning, not a three-stage progressive training curriculum. The candidate focuses on chain-of-thought reasoning for question answering, while the original paper addresses structured visual generation with a three-stage training pipeline (unified alignment, hybrid visual learning, thinking enhancement).

Contribution 3: StructBench benchmark and StructScore evaluation metric

Description: The authors introduce StructBench, a benchmark with over 2,000 samples across six categories for structured image generation and editing. They also propose StructScore, a novel metric that uses VLMs in a multi-round question-answer protocol to evaluate fine-grained factual accuracy and reduce hallucinations compared to naive VLM-as-a-Judge approaches.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. DSG-GAN: Multi-turn text-to-image synthesis via dual semantic-stream guidance with global and local linguistics

URL: View paper

Brief Assessment

DSG-GAN[54] focuses on multi-turn text-to-image synthesis for general visual content manipulation, not structured image generation (charts, diagrams, mathematical figures). It does not present benchmarks or evaluation metrics for structured visuals with multi-round Q&A protocols.

---

2. Gptdrawer: Enhancing visual synthesis through chatgpt

URL: View paper

Brief Assessment

GPTDrawer[51] focuses on enhancing text-to-image generation through prompt refinement using ChatGPT and evaluates outputs using BLIP's cosine similarity metrics. It does not propose a comprehensive benchmark for structured image generation with multi-round Q&A evaluation protocols as described in the original paper's StructBench and StructScore.

---

3. ChatEdit: Towards Multi-turn Interactive Facial Image Editing via Dialogue

URL: View paper

Brief Assessment

ChatEdit[53] focuses on multi-turn interactive facial image editing via dialogue, not structured image generation (charts, diagrams, math figures). The benchmark and evaluation approach are fundamentally different domains.

---

4. What Do You Want? User-centric Prompt Generation for Text-to-image Synthesis via Multi-turn Guidance

URL: View paper

Brief Assessment

User Centric Prompt[52] focuses on user-centric prompt generation for text-to-image synthesis through multi-turn dialogue, not on benchmarking structured image generation with multi-round Q&A evaluation metrics.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Factuality Matters: When Image Generation and Editing Meet Structured Visuals View paper
- [1] Vqgan-clip: Open domain image generation and editing with natural language guidance View paper
- [2] Structured 3D Latents for Scalable and Versatile 3D Generation View paper
- [3] SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion View paper
- [4] Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion View paper
- [5] Unpaired Synthesis of IC Scanning Electron Microscopy Images with Structural Constraints View paper
- [6] DreamOmni: Unified Image Generation and Editing View paper
- [7] Sdedit: Guided image synthesis and editing with stochastic differential equations View paper
- [8] A Comprehensive Survey of Scene Graphs: Generation and Application View paper
- [9] Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN View paper
- [10] Controllable image synthesis methods, applications and challenges: a comprehensive survey View paper
- [11] From Words to Structured Visuals: A Benchmark and Framework for Text-to-Diagram Generation and Editing View paper
- [12] Composer: Creative and Controllable Image Synthesis with Composable Conditions View paper
- [13] MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing View paper
- [14] Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing View paper
- [15] More control for free! image synthesis with semantic diffusion guidance View paper
- [16] PST-Diff: achieving high-consistency stain transfer by diffusion models with pathological and structural constraints View paper
- [17] LayerCraft: Enhancing Text-to-Image Generation with CoT Reasoning and Layered Object Integration View paper
- [18] Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN View paper
- [19] DrawingInStyles: Portrait image generation and editing with spatially conditioned StyleGAN View paper
- [20] High-resolution image synthesis and semantic manipulation with conditional gans View paper
- [21] RAVEL: Rare Concept Generation and Editing via Graph-driven Relational Guidance View paper
- [22] High fidelity text to image generation with contrastive alignment and structural guidance View paper
- [23] CT to ultrasound abdominal image synthesis using domain-constraint generative adversarial networks View paper
- [24] Expressive Image Generation and Editing with Rich Text View paper
- [25] Generative visual manipulation on the natural image manifold View paper
- [26] Controllable data generation by deep learning: A review View paper
- [27] ToddlerDiffusion: Interactive Structured Image Generation with Cascaded Schr\" odinger Bridge View paper
- [28] Deep Image Synthesis, Analysis and Indexing Using Integrated CNN Architectures View paper
- [29] Training-free structured diffusion guidance for compositional text-to-image synthesis View paper
- [30] Pix2video: Video editing using image diffusion View paper
- [31] CREA: A Collaborative Multi-Agent Framework for Creative Image Editing and Generation View paper
- [32] GenPTW: In-Generation Image Watermarking for Provenance Tracing and Tamper Localization View paper
- [33] From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning View paper
- [34] FireFlow: Fast Inversion of Rectified Flow for Image Semantic Editing View paper
- [35] Dreamedit: Subject-driven image editing View paper
- [36] Shape-controlled synthesis of silver and gold nanostructures View paper
- [37] Prompt-to-Prompt Image Editing with Cross Attention Control View paper
- [38] Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing View paper
- [39] Describe, Don't Dictate: Semantic Image Editing with Natural Language Intent View paper
- [40] Sequential attention GAN for interactive image editing View paper
- [41] Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint View paper
- [42] Multi-region text-driven manipulation of diffusion imagery View paper
- [43] Image inpainting based on structural constraint and multi-scale feature fusion View paper
- [44] WeatherEdit: Controllable Weather Editing with 4D Gaussian Field View paper
- [45] Learning hierarchical semantic image manipulation through structured representations View paper
- [46] Image-POSER: Reflective RL for Multi-Expert Image Generation and Editing View paper
- [47] MED-INPAINT: Medical image synthesis using multi-level conditional inpainting with a denoising diffusion probabilistic model and adaptive contrast priors View paper
- [48] Constrained synthesis with projected diffusion models View paper
- [49] A Novel Approach to Industrial Defect Generation through Blended Latent Diffusion Model with Online Adaptation View paper
- [50] MIGC++: Advanced Multi-Instance Generation Controller for Image Synthesis View paper
- [51] Gptdrawer: Enhancing visual synthesis through chatgpt View paper
- [52] What Do You Want? User-centric Prompt Generation for Text-to-image Synthesis via Multi-turn Guidance View paper
- [53] ChatEdit: Towards Multi-turn Interactive Facial Image Editing via Dialogue View paper
- [54] DSG-GAN: Multi-turn text-to-image synthesis via dual semantic-stream guidance with global and local linguistics View paper

- [55] Star: A benchmark for situated reasoning in real-world videos View paper
- [56] Embodiedgpt: Vision-language pre-training via embodied chain of thought View paper
- [57] Chartgemma: Visual instruction-tuning for chart reasoning in the wild View paper
- [58] Measuring and improving chain-of-thought reasoning in vision-language models View paper
- [59] Benchmarking multimodal cot reward model stepwise by visual program View paper
- [60] Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images View paper
- [61] Simple o3: Towards Interleaved Vision-Language Reasoning View paper
- [62] AnnotatedTables: A Large Tabular Dataset with Language Model Annotations View paper
- [63] Translating a visual lego manual to a machine-executable plan View paper
- [64] MINT-CoT: Enabling Interleaved Visual Tokens in Mathematical Chain-of-Thought Reasoning View paper
- [65] Multimodal chain-of-thought reasoning in language models View paper
- [66] Aguvis: Unified pure vision agents for autonomous gui interaction View paper
- [67] Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning View paper
- [68] Reasonrec: A reasoning-augmented multimodal agent for unified recommendation View paper
- [69] Enhancing Spatial Reasoning in Multimodal Large Language Models through Reasoning-based Segmentation View paper
- [70] Mutex: Learning unified policies from multimodal task specifications View paper
- [71] Unified multimodal chain-of-thought reward model through reinforcement fine-tuning View paper
- [72] TinyRS-R1: Compact Multimodal Language Model for Remote Sensing View paper
- [73] Universal Visuo-Tactile Video Understanding for Embodied Interaction View paper
- [74] Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing View paper