Novelty Assessment Report

Paper: ImageDoctor: Diagnosing Text-to-Image Generation via Grounded Image Reasoning
PDF URL: https://openreview.net/pdf?id=04HwYGgp2w
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

The rapid advancement of text-to-image (T2I) models has increased the need for reliable human preference modeling, a demand further amplified by recent progress in reinforcement learning for preference alignment. However, existing approaches typically quantify the quality of a generated image using a single scalar, limiting their ability to provide comprehensive and interpretable feedback on image quality. To address this, we introduce ImageDoctor, a unified multi-aspect T2I model evaluation framework that assesses image quality across four complementary dimensions: plausibility, semantic alignment, aesthetics, and overall quality. ImageDoctor also provides pixel-level flaw indicators in the form of heatmaps, which highlight misaligned or implausible regions, and can be used as a dense reward for T2I model preference alignment. Inspired by the diagnostic process, we improve the detail sensitivity and reasoning capability of ImageDoctor by introducing a ``look-think-predict'' paradigm, where the model first localizes potential flaws, then generates reasoning, and finally concludes the evaluation with quantitative scores. Built on top of a vision-language model and trained through a combination of supervised fine-tuning and reinforcement learning, ImageDoctor demonstrates strong alignment with human preference across multiple datasets, establishing its effectiveness as an evaluation metric. Furthermore, when used as a reward model for preference tuning, ImageDoctor significantly improves generation quality—achieving an improvement of 10% over scalar-based reward models.

Core Task Landscape

This paper addresses: **Multi-aspect Text-to-Image Generation Quality Evaluation with Spatial Localization**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Spatial Control and Layout-Guided Generation**
- **Text-Prompt-Based Spatial Control**
- **Evaluation Frameworks and Benchmarks**
- **Generative Model Architectures and Training Paradigms**
- **Domain-Specific and Application-Oriented Generation**
- **Survey and Comparative Analysis**
- **Auxiliary Generation and Reasoning Tasks**

Complete Taxonomy Tree

- Multi-aspect Text-to-Image Generation Quality Evaluation with Spatial Localization Survey Taxonomy
- Spatial Control and Layout-Guided Generation
  - Box-Constrained and Region-Based Generation (6 papers)
  - [1] Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion (Jinheng Xie, 2023) View paper
  - [10] Eligen: Entity-level controlled image generation with regional attention (Zhang Hong, 2025) View paper
  - [12] Reco: Region-controlled text-to-image generation (Zhengyuan Yang, 2023) View paper
  - [14] Grounding Text-To-Image Diffusion Models For Controlled High-Quality Image Generation (Ahmad SÃ¼leyman, 2025) View paper
  - [19] Migc: Multi-instance generation controller for text-to-image synthesis (Dewei Zhou, 2024) View paper
  - [49] IFAdapter: Instance Feature Control for Grounded Text-to-Image Generation (Wu, 2024) View paper
  - Segmentation-Based Layout Control (4 papers)
  - [5] Spatext: Spatio-textual representation for controllable image generation (Omri Avrahami, 2023) View paper
  - [20] Compositional Text-to-Image Generation with Dense Blob Representations (Nie, 2024) View paper
  - [31] Law-diffusion: Complex scene generation by diffusion with layouts (Binbin Yang, 2023) View paper
  - [41] Layout-Conditioned Autoregressive Text-to-Image Generation via Structured Masking (Zheng, 2025) View paper
  - Sketch-Guided Spatial Control (2 papers)
  - [3] Controllable text-to-image generation with gpt-4 (Zhang, 2023) View paper
  - [44] Sketch-Guided Text-to-Image Generation with Spatial Control (Tianyu Zhang, 2024) View paper
  - Zero-Shot and Training-Free Layout Conditioning (3 papers)
  - [33] Spatial Transport Optimization by Repositioning Attention Map for Training-Free Text-to-Image Synthesis (Lee Yeonkyung, 2025) View paper
  - [38] Check locate rectify: A training-free layout calibration system for text-to-image generation (Biao Gong, 2024) View paper
  - [45] Zero-shot spatial layout conditioning for text-to-image diffusion models (Guillaume Couairon, 2023) View paper
  - LLM-Assisted Layout Generation (2 papers)
  - [37] ComposeAnything: Composite Object Priors for Text-to-Image Generation (Khan, 2025) View paper
  - [42] Llm blueprint: Enabling text-to-image generation with complex and detailed prompts (Gani, 2023) View paper

- Text-Prompt-Based Spatial Control
  - Attention-Based Localization and Refinement (4 papers)
  - [6] Grounded Text-to-Image Synthesis with Attention Refocusing (Quynh Phung, 2023) View paper
  - [7] Localized text-to-image generation for free via cross attention control (He Yutong, 2023) View paper
  - [40] Semantic Attention and LLM-based Layout Guidance for Text-to-Image Generation (Yuxiang Song, 2025) View paper
  - [43] Semantic-Spatial Attention for Refined Object Placement in Text-to-Image Synthesis (Jianwei Zheng, 2025) View paper
  - Hierarchical Cross-Modal Alignment (1 papers)
  - [46] HCMA: Hierarchical Cross-model Alignment for Grounded Text-to-Image Generation (Wang Hang, 2025) View paper
  - Spatial Relation Dataset Construction and Fine-Tuning (3 papers)
  - [30] Improving Explicit Spatial Relationships in Text-to-Image Generation through an Automatically Derived Dataset (Salaberria, 2024) View paper
  - [34] Compass: Enhancing spatial understanding in text-to-image diffusion models (Zhang Gaoyang, 2025) View paper
  - [50] SELMA: Learning and Merging Skill-Specific Text-to-Image Experts with Auto-Generated Data (Li, 2024) View paper
- Evaluation Frameworks and Benchmarks
  - Multi-Aspect Quality Assessment ★ (2 papers)
  - [0] ImageDoctor: Diagnosing Text-to-Image Generation via Grounded Image Reasoning (Anon et al., 2026) View paper
  - [9] Quality assessment for text-to-image generation: A survey (Yu Tian, 2025) View paper
  - Compositional and Spatial Relation Benchmarks (6 papers)
  - [8] GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment (Ghosh, 2023) View paper
  - [15] T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation (Kaiyi Huang, 2023) View paper
  - [18] T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation (Kaiyi Huang, 2023) View paper
  - [22] Evaluating the Generation of Spatial Relations in Text and Image Generative Models (Lee, 2024) View paper
  - [35] Draw ALL Your Imagine: A Holistic Benchmark and Agent Framework for Complex Instruction-based Image Generation (Zhou Yucheng, 2025) View paper
  - [39] Evaluating a text-to-image modelâ□□s understanding of spatial relations and spatial prepositions (H, 2023) View paper
  - VQA-Based Faithfulness Evaluation (2 papers)
  - [4] TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering (Yushi Hu, 2023) View paper
  - [28] Tokenfocus-VQA: Enhancing Text-to-Image Alignment with Position-Aware Focus and Multi-Perspective Aggregations on LVLMs (Zhang Zijian, 2025) View paper
  - World Knowledge and Semantic Evaluation (2 papers)
  - [16] Wise: A world knowledge-informed semantic evaluation for text-to-image generation (Niu Yuwei, 2025) View paper
  - [25] DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models (Jaemin Cho, 2022) View paper
- Generative Model Architectures and Training Paradigms
  - Masked Generative Transformers (1 papers)
  - [2] Muse: Text-To-Image Generation via Masked Generative Transformers (Chang, 2023) View paper
  - Diffusion Model Variants and Enhancements (1 papers)
  - [32] RSVQ-Diffusion Model for Text-to-Remote-Sensing Image Generation (Xin Gao, 2025) View paper
  - Multiview and 3D-Aware Generation (1 papers)
  - [36] CoSER: Towards Consistent Dense Multiview Text-To-Image Generator for 3D Creation (Bonan Li, 2025) View paper
- Domain-Specific and Application-Oriented Generation
  - Design and Typography Generation (1 papers)
  - [27] Designdiffusion: High-quality text-to-design image generation with diffusion models (Wang, 2025) View paper
  - Remote Sensing Image Synthesis (1 papers)
  - [21] Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework (Junyan Ye, 2025) View paper
  - Object-Level Control and Exploration (3 papers)
  - [17] FOCUS: Unified Vision-Language Modeling for Interactive Editing Driven by Referential Segmentation (Yang Fan, 2025) View paper
  - [23] Localizing object-level shape variations with text-to-image diffusion models (Or Patashnik, 2023) View paper
  - [26] VODiff: Controlling Object Visibility Order in Text-to-Image Generation (Dong Liang, 2025) View paper
- Survey and Comparative Analysis (3 papers)
  - [13] Text-to-Image Synthesis: A Decade Survey (Zhang, 2024) View paper
  - [24] Explainable Image-Centric Forgery Detection: A Survey (Xinyu Wu, 2025) View paper
  - [47] A comparative analysis and investigation of Attn-GAN and SSA-GAN for text-to-image generation (Zhang, 2024) View paper
- Auxiliary Generation and Reasoning Tasks (3 papers)
  - [11] End-to-end text-to-image synthesis with spatial constrains (Min Wang, 2020) View paper
  - [29] Visual programming for step-by-step text-to-image generation and evaluation (J Cho, 2023) View paper
  - [48] Getting it Right: Improving Spatial Consistency in Text-to-Image Models (Stan, 2024) View paper

Narrative

Core task: Multi-aspect text-to-image generation quality evaluation with spatial localization. The field has evolved around several interconnected branches that address both generation and assessment challenges. Spatial Control and Layout-Guided Generation encompasses methods that use explicit spatial constraints—such as bounding boxes (Boxdiff[1]), attention mechanisms (Attention Refocusing[6], Localized Cross Attention[7]), or layout priors (Zero-shot Layout[45])—to guide where objects appear in synthesized images. Text-Prompt-Based Spatial Control explores how natural language descriptions can encode spatial relationships (Spatext[5], Spatial Prepositions[39]) without requiring structured annotations. Meanwhile, Generative Model Architectures and Training Paradigms investigates the underlying diffusion and transformer frameworks (Muse[2], Grounding Diffusion[14]) that enable fine-grained control. Domain-Specific and Application-Oriented Generation targets specialized contexts like satellite imagery (Satellite Street View[21]) or design tasks (DesignDiffusion[27]), while Auxiliary Generation and Reasoning Tasks address complementary problems such as visual reasoning and compositional understanding.

Evaluation Frameworks and Benchmarks have become central to measuring how well models satisfy complex prompts across multiple quality dimensions—attribute binding, spatial accuracy, object counting, and compositional fidelity. Works like TIFA[4], GenEval[8], and T2I-CompBench[18] introduced systematic test suites, while Quality Assessment Survey[9] and Decade Survey[13] provide broader perspectives on progress and open challenges. Within this evaluation landscape, ImageDoctor[0] emphasizes multi-aspect quality assessment with spatial localization, positioning itself alongside efforts that diagnose specific failure modes and provide interpretable feedback. Compared to holistic benchmarks (T2I-CompBench++[15], DALL-EVAL[25]) that aggregate scores across many prompts, ImageDoctor[0] focuses on pinpointing where and why generation quality degrades, offering a more granular diagnostic lens. This contrasts with purely compositional metrics (GenEval[8]) or attribute-binding tests (TIFA[4]), highlighting a shift toward spatially aware, interpretable evaluation that can guide iterative model improvement.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

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

Authors: Yu Tian, Yue Liu, Shiqi Wang, Sam Kwong | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ The technical prior branch uses a detection model to localize possible perceptual artifacts â¦ the spatial relationship between two objects and evaluate the score of the spatial relationship â¦

Relationship Analysis

Both papers belong to the Multi-Aspect Quality Assessment category, focusing on evaluation frameworks that assess multiple quality dimensions of text-to-image generation with spatial localization capabilities. The candidate paper appears to be a survey that reviews quality assessment methods including artifact detection and spatial relationship evaluation, while the original paper (ImageDoctor) presents a specific unified evaluation framework with a novel 'look-think-predict' paradigm that generates four-dimensional scores, heatmaps for flaw localization, and interpretable reasoning chains. The key difference is that the original paper proposes a concrete system with grounded image reasoning and reinforcement learning training, whereas the candidate is a survey reviewing the broader landscape of T2I quality assessment approaches.

Contributions Analysis

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

Contribution 1: ImageDoctor unified multi-aspect T2I evaluation framework

Description: The authors propose ImageDoctor, a unified framework that evaluates text-to-image generation across four dimensions (plausibility, semantic alignment, aesthetics, overall quality) and provides pixel-level flaw indicators as heatmaps highlighting misaligned or implausible regions.

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. A survey on quality metrics for text-to-image generation

URL: View paper

Brief Assessment

Quality Metrics Survey[56] provides a comprehensive taxonomy of T2I evaluation metrics but does not propose a specific unified framework with multi-aspect scoring and pixel-level heatmaps like ImageDoctor. The survey categorizes existing metrics rather than introducing a novel evaluation system with the 'look-think-predict' paradigm and dense spatial feedback.

---

2. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models

URL: View paper

Brief Assessment

HRS-Bench[59] focuses on benchmarking T2I models across 13 skills using detection-based and alignment-based metrics, rather than providing a unified evaluation framework with pixel-level diagnostic feedback like ImageDoctor's heatmaps and reasoning chains.

---

3. Holistic evaluation of text-to-image models

URL: View paper

Brief Assessment

Holistic Evaluation[53] focuses on evaluating existing T2I models across 12 aspects using standardized scenarios and metrics, whereas ImageDoctor proposes a novel unified model architecture that generates both multi-dimensional scores and pixel-level heatmaps with grounded reasoning capabilities. The candidate is an evaluation benchmark, not a unified evaluation model framework.

---

4. Evaluating text-to-visual generation with image-to-text generation

URL: View paper

Brief Assessment

Image-to-Text Evaluation[54] focuses on text-to-visual alignment using VQA-based scoring (VQAScore) rather than providing a unified multi-aspect evaluation framework with pixel-level heatmaps. The candidate does not address plausibility assessment, aesthetic evaluation, or spatial flaw localization through heatmaps.

---

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

URL: View paper

Brief Assessment

Quality Assessment Survey[9] is a survey paper that reviews existing evaluation methods for text-to-image generation. It does not propose a unified framework with multi-aspect scoring and pixel-level heatmaps like ImageDoctor.

---

6. Visual programming for step-by-step text-to-image generation and evaluation

URL: View paper

Brief Assessment

Visual Programming[29] focuses on visual programming frameworks for T2I generation and evaluation using modular visual programs, while ImageDoctor proposes a unified MLLM-based framework with pixel-level heatmaps and multi-dimensional scoring. The candidate's evaluation approach uses separate expert modules (object detection, OCR, VQA) combined via programs, whereas ImageDoctor integrates evaluation within a single model architecture with a heatmap decoder.

---

7. UniGenBench++: A Unified Semantic Evaluation Benchmark for Text-to-Image Generation

URL: View paper

Brief Assessment

UniGenBench++[58] focuses on benchmark construction for semantic evaluation across diverse prompt themes and languages, not on developing a unified evaluation framework with pixel-level diagnostics and dense reward signals like ImageDoctor.

---

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

URL: View paper

Brief Assessment

Human Preference v2[57] focuses on training a preference prediction model (HPS v2) that outputs scalar scores for human preference alignment, not a unified multi-aspect framework with pixel-level heatmaps and grounded reasoning as proposed in ImageDoctor.

---

9. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation

URL: View paper

Brief Assessment

T2I-CompBench[18] focuses on compositional text-to-image generation evaluation (attribute binding, object relationships, complex compositions) rather than providing a unified multi-aspect framework with pixel-level flaw indicators and heatmaps as ImageDoctor does.

---

10. Multimodal Benchmarking and Recommendation of Text-to-Image Generation Models

URL: View paper

Brief Assessment

Multimodal Benchmarking[55] focuses on quantitative metrics (CLIP similarity, LPIPS, FID) and metadata-augmented prompts for model selection, not on providing interpretable multi-dimensional scores with pixel-level flaw localization heatmaps as ImageDoctor does.

Contribution 2: Look-think-predict paradigm for grounded image reasoning

Description: The authors introduce a diagnostic reasoning paradigm where the model first localizes potential flaw regions (look), analyzes them through structured reasoning (think), and then produces final evaluation scores and heatmaps (predict), mimicking human evaluation processes.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. VGR: Visual Grounded Reasoning

URL: View paper

Brief Assessment

VGR[51] focuses on visual grounded reasoning where models detect regions first then reason, while the original paper's look-think-predict paradigm is specifically designed for text-to-image evaluation with flaw localization. VGR[51] addresses multimodal comprehension tasks, not T2I quality assessment.

---

2. A computational model of event segmentation from perceptual prediction

URL: View paper

Brief Assessment

Event Segmentation[52] focuses on event segmentation from perceptual prediction in cognitive modeling, not image quality evaluation or text-to-image generation diagnostics.

Contribution 3: DenseFlow-GRPO reinforcement learning framework

Description: The authors present DenseFlow-GRPO, a novel reinforcement learning method that incorporates both image-level and pixel-level dense reward signals from ImageDoctor to provide spatially aligned supervision for text-to-image model training, enabling fine-grained region-aware optimization.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Aligning Text-to-Image Diffusion Models With Constrained Reinforcement Learning

URL: View paper

Brief Assessment

Constrained RL[63] focuses on temporal policy optimization with step-specific rewards and constrained RL for diffusion models, not on incorporating dense pixel-level spatial rewards from heatmaps for region-aware text-to-image optimization.

---

2. Subject-driven Text-to-Image Generation via Preference-based Reinforcement Learning

URL: View paper

Brief Assessment

Subject-driven Preference[68] focuses on subject-driven text-to-image generation using preference-based RL with a λ-harmonic reward function, not on dense pixel-level rewards for general T2I training. The candidate does not address spatially grounded heatmap-based dense rewards as proposed in the original paper's DenseFlow-GRPO framework.

---

3. Step-level Reward for Free in RL-based T2I Diffusion Model Fine-tuning

URL: View paper

Brief Assessment

Step-level Reward[67] focuses on credit assignment via cosine similarity-based reward redistribution across denoising timesteps, not on incorporating pixel-level dense spatial rewards from heatmaps for region-aware optimization as in DenseFlow-GRPO.

---

4. Designing, learning, and utilizing dense reward for generative AI alignment

URL: View paper

Brief Assessment

Dense Reward Design[65] discusses dense reward principles for text-to-image diffusion models but does not present a specific framework incorporating pixel-level heatmap rewards like DenseFlow-GRPO.

---

5. Pixel-wise RL on Diffusion Models: Reinforcement Learning from Rich Feedback

URL: View paper

Prior Art Analysis

Pixel-wise RL[64] demonstrates that pixel-level dense reward signals for diffusion model training were already proposed and implemented. The candidate paper presents the pixel-wise policy optimisation (pxpo) algorithm that provides pixel-wise feedback through heatmaps to diffusion models, enabling fine-grained spatial supervision. This directly challenges the novelty claim of DenseFlow-GRPO, as pxpo already incorporated both image-level and pixel-level dense rewards for spatially aligned optimization of generative models before the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers propose reinforcement learning frameworks that incorporate pixel-wise dense rewards for diffusion model training. The candidate's pxpo algorithm already provides spatially aligned supervision through pixel-wise feedback heatmaps, which is the core innovation claimed by DenseFlow-GRPO. - **Original**: we introduce denseflow-grpo, a new rlhf framework that enhances t2i models with both image-level and pixel-level dense reward signals. by leveraging the rich diagnostic feedback from imagedoctor, denseflow-grpo delivers more precise and spatially aligned supervision, enabling t2i models to learn not... - **Candidate**: we introduce the pixel-wise policy optimisation (pxpo) algorithm, a technique that allows ddim models to receive pixel-wise feedback from a black-box function that produces a single-channel heatmap. we show that pxpo can generalise from a small sample size without needing to train a reward model.

---

6. Visual-CoG: Stage-Aware Reinforcement Learning with Chain of Guidance for Text-to-Image Generation

URL: View paper

Brief Assessment

Visual-CoG[60] focuses on stage-aware RL for autoregressive text-to-image models with chain-of-thought reasoning, not on dense pixel-level rewards from spatial heatmaps for flow-based diffusion models.

---

7. Seeing What Matters: Visual Preference Policy Optimization for Visual Generation

URL: View paper

Prior Art Analysis

Visual Preference[66] demonstrates that prior work exists on incorporating dense, pixel-level rewards into flow-based reinforcement learning frameworks for visual generation. Both papers reformulate flow matching into SDE for RL training and introduce pixel-level advantage mechanisms. Visual Preference[66] presents 'visual preference policy optimization (vipo)' which lifts scalar feedback into structured, pixel-level advantages using a perceptual structuring module, predating the original paper's DenseFlow-GRPO approach. The candidate explicitly describes converting coarse scalar advantages into spatially-resolved advantages through allocation maps, providing fine-grained supervision similar to the original paper's dense reward formulation.

Evidence

Evidence 1 - **Rationale**: Both papers describe novel RL methods that extend GRPO with pixel-level/spatial feedback mechanisms for visual generation, demonstrating that the concept of dense, spatially-aware rewards in flow-based RL existed prior to the original submission. - **Original**: we present denseflow-grpo, a novel t2i reinforcement learning method that incorporates imagedoctor's dense spatial feedback into the reward signal, providing region-aware supervision and leading to more robust improvements in image generation. - **Candidate**: we introduce visual preference policy optimization (vipo), a grpo variant that lifts scalar feedback into structured, pixel-level advantages. vipo employs a perceptual structuring module that uses pretrained vision backbones to construct spatially and temporally aware advantage maps, redistributing ...

Evidence 2 - **Rationale**: Both papers identify the same limitation in existing GRPO approaches (sparse image-level rewards) and propose the same solution direction (incorporating pixel-level/dense spatial feedback), indicating that Visual Preference[66] addressed this problem space first. - **Original**: however, current rlhf approaches, such as flow-grpo (liu et al., 2025), rely solely on sparseimage-levelrewards, which overlook spatially localized feedback and thus fail to provide fine-grained guidance during training. to address this limitation, we introducedenseflow-grpo, a new rlhf framework th... - **Candidate**: however, existing grpo pipelines rely on a single scalar reward per sample, treating each image or video as a holistic entity and ignoring the rich spatial and temporal structure of visual content. this coarse supervision hinders the correction of localized artifacts and the modeling of fine-grained...

Evidence 3 - **Rationale**: Both papers describe frameworks that provide spatially aligned, fine-grained supervision for visual generation through region-aware optimization, demonstrating conceptual overlap in the core contribution. - **Original**: By leveraging the rich diagnostic feedback from imagedoctor, denseflow-grpo delivers more precise and spatially aligned supervision, enabling t2i models to learn not only what constitutes a good image globally, but also how to refine local regions in a fine-grained manner. - **Candidate**: This design allows vipo to emphasize visually informative regions, yielding fine-grained alignment with perceptual preferences while maintaining the stability and simplicity of the original grpo algorithm. in this section, we first present the preliminaries of applying grpo to visual generation, and...

---

8. Listener-Rewarded Thinking in VLMs for Image Preferences

URL: View paper

Brief Assessment

Listener-Rewarded Thinking[62] focuses on improving reward models for human preference alignment through listener-augmented GRPO, not on dense pixel-level rewards for text-to-image training. The candidate addresses reasoning accuracy in reward models, while the original contribution specifically targets spatially aligned supervision with pixel-level dense rewards from ImageDoctor heatmaps for fine-grained region-aware optimization in T2I generation.

---

9. A dense reward view on aligning text-to-image diffusion with preference

URL: View paper

Brief Assessment

Dense Reward[61] focuses on introducing temporal discounting into DPO-style objectives for text-to-image diffusion models, emphasizing initial steps of the reverse chain. The original paper's DenseFlow-GRPO uses pixel-level dense rewards from ImageDoctor heatmaps for spatially aligned supervision, which is a distinct approach not addressed in Dense Reward[61].

Appendix: Text Similarity Detection

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

References

- [0] ImageDoctor: Diagnosing Text-to-Image Generation via Grounded Image Reasoning View paper
- [1] Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion View paper

- [2] Muse: Text-To-Image Generation via Masked Generative Transformers View paper
- [3] Controllable text-to-image generation with gpt-4 View paper
- [4] TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering View paper
- [5] Spatext: Spatio-textual representation for controllable image generation View paper
- [6] Grounded Text-to-Image Synthesis with Attention Refocusing View paper
- [7] Localized text-to-image generation for free via cross attention control View paper
- [8] GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment View paper
- [9] Quality assessment for text-to-image generation: A survey View paper
- [10] Eligen: Entity-level controlled image generation with regional attention View paper
- [11] End-to-end text-to-image synthesis with spatial constrains View paper
- [12] Reco: Region-controlled text-to-image generation View paper
- [13] Text-to-Image Synthesis: A Decade Survey View paper
- [14] Grounding Text-To-Image Diffusion Models For Controlled High-Quality Image Generation View paper
- [15] T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation View paper
- [16] Wise: A world knowledge-informed semantic evaluation for text-to-image generation View paper
- [17] FOCUS: Unified Vision-Language Modeling for Interactive Editing Driven by Referential Segmentation View paper
- [18] T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation View paper
- [19] Migc: Multi-instance generation controller for text-to-image synthesis View paper
- [20] Compositional Text-to-Image Generation with Dense Blob Representations View paper
- [21] Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework View paper
- [22] Evaluating the Generation of Spatial Relations in Text and Image Generative Models View paper
- [23] Localizing object-level shape variations with text-to-image diffusion models View paper
- [24] Explainable Image-Centric Forgery Detection: A Survey View paper
- [25] DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models View paper
- [26] VODiff: Controlling Object Visibility Order in Text-to-Image Generation View paper
- [27] Designdiffusion: High-quality text-to-design image generation with diffusion models View paper
- [28] Tokenfocus-VQA: Enhancing Text-to-Image Alignment with Position-Aware Focus and Multi-Perspective Aggregations on LVLMs View paper
- [29] Visual programming for step-by-step text-to-image generation and evaluation View paper
- [30] Improving Explicit Spatial Relationships in Text-to-Image Generation through an Automatically Derived Dataset View paper
- [31] Law-diffusion: Complex scene generation by diffusion with layouts View paper
- [32] RSVQ-Diffusion Model for Text-to-Remote-Sensing Image Generation View paper
- [33] Spatial Transport Optimization by Repositioning Attention Map for Training-Free Text-to-Image Synthesis View paper
- [34] Compass: Enhancing spatial understanding in text-to-image diffusion models View paper
- [35] Draw ALL Your Imagine: A Holistic Benchmark and Agent Framework for Complex Instruction-based Image Generation View paper
- [36] CoSER: Towards Consistent Dense Multiview Text-To-Image Generator for 3D Creation View paper
- [37] ComposeAnything: Composite Object Priors for Text-to-Image Generation View paper
- [38] Check locate rectify: A training-free layout calibration system for text-to-image generation View paper
- [39] Evaluating a text-to-image modelâ□□s understanding of spatial relations and spatial prepositions View paper
- [40] Semantic Attention and LLM-based Layout Guidance for Text-to-Image Generation View paper
- [41] Layout-Conditioned Autoregressive Text-to-Image Generation via Structured Masking View paper
- [42] Llm blueprint: Enabling text-to-image generation with complex and detailed prompts View paper
- [43] Semantic-Spatial Attention for Refined Object Placement in Text-to-Image Synthesis View paper
- [44] Sketch-Guided Text-to-Image Generation with Spatial Control View paper
- [45] Zero-shot spatial layout conditioning for text-to-image diffusion models View paper
- [46] HCMA: Hierarchical Cross-model Alignment for Grounded Text-to-Image Generation View paper
- [47] A comparative analysis and investigation of Attn-GAN and SSA-GAN for text-to-image generation View paper
- [48] Getting it Right: Improving Spatial Consistency in Text-to-Image Models View paper
- [49] IFAdapter: Instance Feature Control for Grounded Text-to-Image Generation View paper
- [50] SELMA: Learning and Merging Skill-Specific Text-to-Image Experts with Auto-Generated Data View paper
- [51] VGR: Visual Grounded Reasoning View paper
- [52] A computational model of event segmentation from perceptual prediction View paper
- [53] Holistic evaluation of text-to-image models View paper
- [54] Evaluating text-to-visual generation with image-to-text generation View paper
- [55] Multimodal Benchmarking and Recommendation of Text-to-Image Generation Models View paper
- [56] A survey on quality metrics for text-to-image generation View paper
- [57] Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis View paper
- [58] UniGenBench++: A Unified Semantic Evaluation Benchmark for Text-to-Image Generation View paper
- [59] Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models View paper
- [60] Visual-CoG: Stage-Aware Reinforcement Learning with Chain of Guidance for Text-to-Image Generation View paper
- [61] A dense reward view on aligning text-to-image diffusion with preference View paper
- [62] Listener-Rewarded Thinking in VLMs for Image Preferences View paper
- [63] Aligning Text-to-Image Diffusion Models With Constrained Reinforcement Learning View paper
- [64] Pixel-wise RL on Diffusion Models: Reinforcement Learning from Rich Feedback View paper
- [65] Designing, learning, and utilizing dense reward for generative AI alignment View paper
- [66] Seeing What Matters: Visual Preference Policy Optimization for Visual Generation View paper
- [67] Step-level Reward for Free in RL-based T2I Diffusion Model Fine-tuning View paper
- [68] Subject-driven Text-to-Image Generation via Preference-based Reinforcement Learning View paper