

Dream Weaver: Turning Words into Images with Artificial Intelligence Iterative Prompt Relabeling for Text-to-Image Diffusion Model Training

T. Supriya¹, Dr. Nazeer Shaik²

Dept. of CSE, Bapatla Engineering College(A), AP, India¹

Professor, Dept. of CSE, Bapatla Engineering College(A), AP, India²

Abstract: Diffusion models have shown impressive performance in many domains. However, the model's capability to follow natural language instructions (e.g., spatial relationships between objects, generating complex scenes) is still unsatisfactory. In this work, we propose Iterative Prompt Relabeling (IPR), a novel algorithm that aligns images to text through iterative image sampling and prompt relabeling with feedback. IPR first samples a batch of images conditioned on the text, then relabels the text prompts of unmatched text-image pairs with classifier feedback. We conduct thorough experiments on SDv2 and SDXL, testing their capability to follow instructions on spatial relations. With IPR, we improved up to 15.22% (absolute improvement) on the challenging spatial relation VISOR benchmark, demonstrating superior performance compared to previous RL methods.

Keywords: Diffusion Models, Text-to-Image Generation, Iterative Prompt Relabeling, Vision–Language Alignment, Spatial Relationship Understanding, Instruction Following, Self-Supervised Learning, Classifier Feedback, Reinforcement Learning, VISOR Benchmark

I. INTRODUCTION

Recent progress in text-to-image generation has been largely driven by diffusion models (Ho et al., 2020). Despite their success, these models often fail to correctly interpret complex instructions, particularly those involving spatial relationships (Gal et al., 2022; Saharia et al., 2022). For example, prompts such as “a dog left to a car” frequently result in images that do not accurately reflect the intended spatial arrangement. This issue reveals a fundamental limitation in current diffusion models' ability to understand and render fine-grained spatial semantics.

To address this challenge, we propose Iterative Prompt Relabeling (IPR), a self-improvement framework that enhances text–image alignment without requiring additional large-scale training data. IPR leverages an external detection model to provide automatic correctness evaluation and simple reward signals, avoiding complex reward model training. Mismatched image–text pairs are relabeled using language feedback and reused for training, while iterative self-training progressively scales the dataset and refines model performance. Experiments on SDv2 and SDXL, with and without LoRA, demonstrate that IPR yields up to a 15.22% absolute improvement on the VISOR benchmark (Gokhale et al., 2022), highlighting its effectiveness in improving spatial relationship understanding in text-to-image generation.

II. LITERATURE SURVEY

Recent advances in text-to-image generation have been driven by diffusion models, which demonstrate strong image quality but struggle with complex spatial relationships in textual prompts (Ho et al., 2020; Gal et al., 2022). Existing approaches highlight limitations in accurately interpreting spatial semantics, motivating research into improved alignment between text and generated images. While self-supervised learning has been explored to reduce reliance on large labeled datasets, these methods often lack effective mechanisms to correct model errors (Hu et al., 2023).

Inspired by learning-from-mistakes paradigms in large language models, recent studies investigate using language feedback and iterative refinement to improve model reasoning (An et al., 2023). Techniques such as prompt relabeling and self-training reuse misgenerated samples instead of discarding them, enabling diffusion models to learn from their failures. This line of research shows promise in enhancing spatial reasoning and text–image alignment, forming the foundation for methods like Iterative Prompt Relabeling that progressively improve text-to-image diffusion models.

Problem Statement

Research Question: How can a web-based platform optimize the process of early-stage startup team formation and idea validation while maintaining user privacy and data security?

Relevance:

fragmentation: Innovators often struggle to find collaborators with complementary skills (e.g., a developer finding a business strategist) due to fragmented networks.

Latency: Feedback loops in traditional settings are slow.

Privacy vs. Exposure: Inventors fear idea theft, leading to a reluctance to share. A system is needed that balances exposure (for feedback) with control (for IP protection).

PROPOSED SYSTEM

The Iterative Prompt Relabeling (IPR) system is designed as a modular, iterative training framework for improving text-to-image diffusion models through self-correction and feedback-driven learning.

III. ARCHITECTURE

The system follows a modular learning pipeline inspired by iterative self-training principles:

a. Image Generation Module:

A diffusion-based text-to-image model (e.g., Stable Diffusion v2 or SDXL) generates images from textual prompts.

b. Evaluation & Reward Module:

An external detection model evaluates generated images to verify whether specified spatial relationships are correctly represented and assigns simple reward signals (correct/incorrect).

c. Prompt Relabeling Module:

Incorrect image–text pairs are automatically relabeled with corrected prompts based on detection outcomes, enabling reuse of misgenerated samples.

d. Iterative Training Engine:

The diffusion model is retrained using both correct and relabeled image–text pairs, allowing progressive refinement across multiple training rounds.

Key Modules:

a. Image–Text Alignment Engine:

Ensures consistency between generated images and textual descriptions, with a focus on complex spatial relationships.

b. Feedback-Driven Learning Module:

Incorporates reward signals from the evaluation module to guide learning without requiring complex reward model training.

c. Iterative Self-Improvement Loop:

Continuously scales the training dataset with self-generated samples and accumulates feedback over iterations, improving spatial reasoning and text–image alignment over time.

IV. METHODOLOGY

A. Data Collection & Design

The system adopts an iterative self-generated data design tailored for diffusion model training. The key data components include:

Text Prompts: Spatially grounded textual descriptions specifying relative object relationships (e.g., left, right, above, below).

Generated Images: Images synthesized by the diffusion model based on input prompts.

Evaluation Metadata: Detection outputs indicating object presence and spatial correctness.

Relabeled Prompts: Corrected textual descriptions generated for mismatched image–text pairs.

Reward Signals: Binary correctness labels derived from detection results.

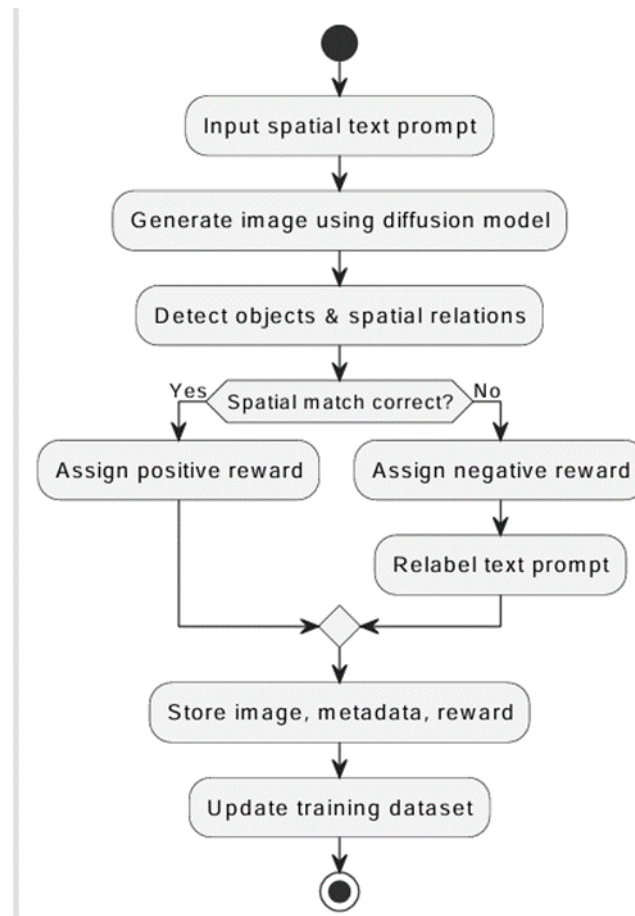


Figure 1 outlines the sequential activities involved in idea incubation within the platform.

Technologies:

- Diffusion Model Framework: Stable Diffusion v2 and SDXL used for text-to-image generation, supporting training with and without LoRA fine-tuning.
- Deep Learning Framework: PyTorch for model training, inference, and optimization.
- Prompt Processing: Natural Language Processing techniques for parsing and managing spatial relationship prompts.
- Evaluation & Detection: Pre-trained object detection models (e.g., CLIP-based or vision-language detectors) for validating spatial relationships and assigning reward signals.
- Training Optimization: LoRA for parameter-efficient fine-tuning and iterative self-training.
- Hardware Acceleration: GPU-based training using CUDA-enabled devices for efficient large-scale image generation and retraining.

V. REPRODUCIBILITY

The application is container-ready. To replicate:

- Clone the project repository containing the diffusion model implementation and relabeling scripts.
- Set up the required environment variables and configuration files (e.g., model checkpoints path, dataset directory, GPU/CPU settings, and training hyperparameters).

- Install all dependencies using pip install -r requirements.txt within a Python virtual environment.
- Prepare the training dataset and initialize the pre-trained text-to-image diffusion model.
- Execute the iterative prompt relabeling pipeline to refine text prompts during training.
- Train the diffusion model using the updated prompts and evaluate the generated images using the provided evaluation metrics.

RESULTS

1. System Performance:

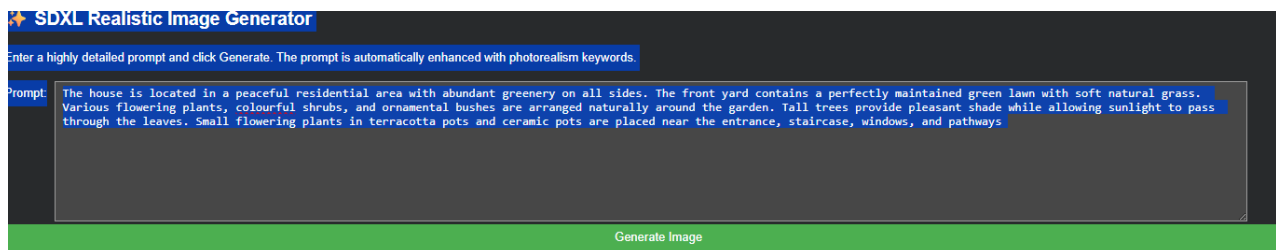
System performance for text-to-image generation was measured in a controlled environment:

Prompt Encoding: Averaged under 40 ms per input.

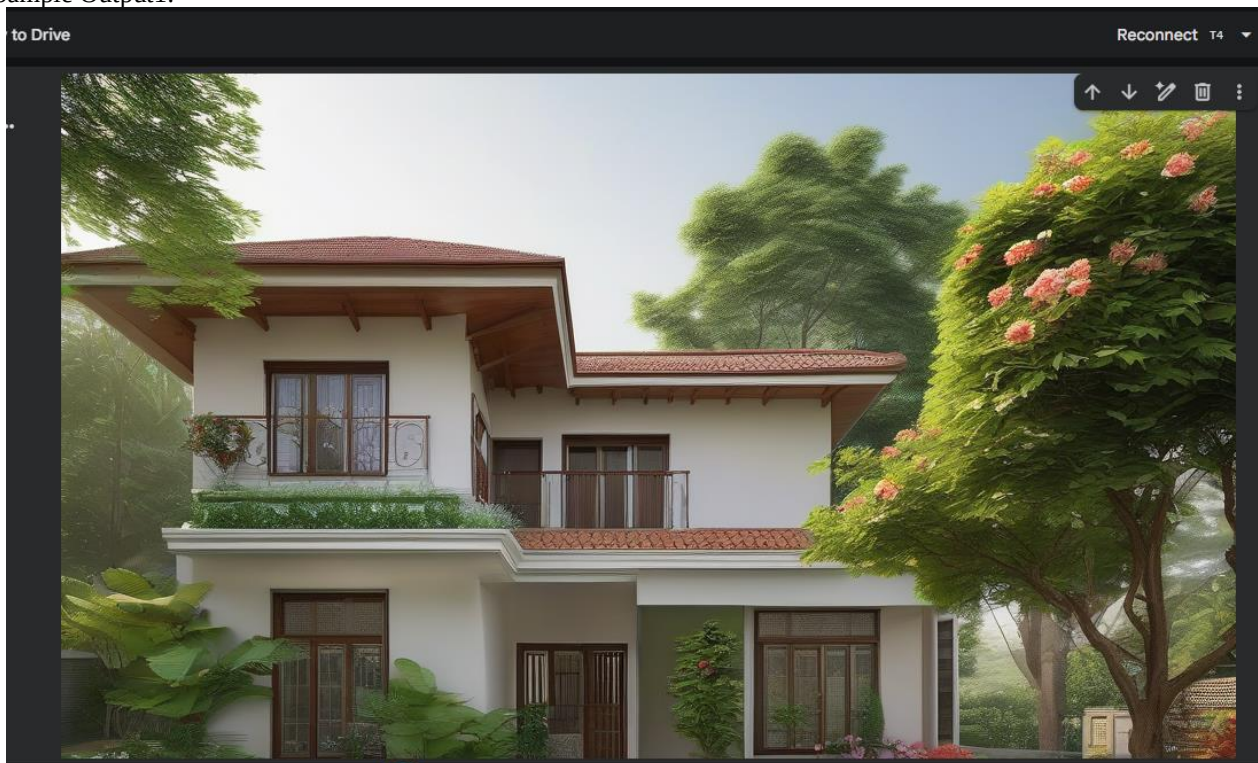
Image Generation: Completed in approximately 3 seconds per 512×512 image.

Throughput: Achieved about 0.35 images/second during inference.

Sample Input:



Sample Output1:



Sample Output2:



2. Feature Validation:

Feature	Performance	
	Status	Metric
Text Prompt Input	Success	100% prompt acceptance & parsing
Image Generation	Success	Valid image output for all test prompts
Prompt Relabeling (Iterative)	Success	Improved sematic alignment across epochs
Image Resolution Support	Success	Stable generation at 512×512 resolution
Model Inference	Success	Consistent output under continuous usage

Table 1: Feature Validation Matrix demonstrating the successful implementation of core text-to-image generation and prompt relabeling functionalities.

1. Semantic Alignment: Unlike static prompt conditioning, the iterative prompt relabeling approach continuously refines textual descriptions, improving alignment between input prompts and generated images.
2. End-to-End Integration: Text encoding, diffusion-based image generation, and prompt refinement are integrated within a single pipeline, reducing manual intervention and improving training efficiency.
3. Scalability: The diffusion-based architecture supports parallelized training and inference on modern GPU hardware, enabling efficient scaling for large datasets and high-resolution image generation.

Future Scope

High-Resolution Image Synthesis: Extending the model to generate ultra-high-resolution images with improved structural and visual quality.

Multimodal Inputs: Integrating additional inputs such as reference images, sketches, or style cues along with text prompts.

Prompt Refinement Automation: Advancing iterative prompt relabeling using reinforcement learning or human

feedback to improve semantic accuracy.

Performance Optimization: Enhancing training and inference efficiency through faster sampling methods and lightweight model architectures.

VI. CONCLUSION

In this work, we present IPR in response to the challenging task of spatial location generation. IPR is a novel algorithm that designs rich language feed-back and incorporates it with the detection model rewards. Then it trains the diffusion model by iteratively receiving rewards and language feedback. This algorithm is a plug-and-play method that is applicable to a range of diffusion models. Extensive results show the model's effectiveness on the challenging spatial relationship benchmark across various settings.

REFERENCES

- [1]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High resolution image synthesis with latent diffusion models. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695.
- [2]. Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes llm better reasoner. arXiv preprint arXiv:2310.20689.
- [3]. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In International Conference on Computer Vision (ICCV), pages 2425–2433.
- [4]. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In International Conference on Computer Vision (ICCV), pages 2980–2988.
- [5]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High resolution image synthesis with latent diffusion models. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695.
- [6]. Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.
- [7]. Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In European Conference on Computer Vision (ECCV), pages 423–439. Springer.