॥ सा विद्या या विमुक्तये ॥
ಭಾ.ತಂ.ಸಂ. ಧಾರವಾಡ
भा. प्रौ. सं. धारवाड
IIT DHARWAD

ಭಾರತೀಯ ತಂತ್ರಜ್ಞಾನ ಸಂಸ್ಥೆ ಧಾರವಾಡ
भारतीय प्रौद्योगिकी संस्थान धारवाड
INDIAN INSTITUTE OF TECHNOLOGY DHARWAD

Mastering Text-to-Image Diffusion Using RPG with Multimodal LLMs

Guide: Prof. Dr. Vandana Bharti

August 5, 2025

Index

1. Introduction
2. Results
3. Inference

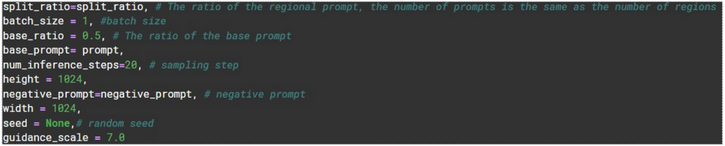
Introduction

- The motivation of the research was to Analyze the behaviour of Different Multimodal Diffusion Models.
- Existing diffusion-based text-to-image models struggle with complex prompts involving multiple objects, attributes, and relationships, limiting their compositional generation abilities.
- The proposed RPG (Recaptioning, Planning, and Generation) framework leverages multimodal LLMs as global planners and introduces complementary regional diffusion for training-free, region-wise text-to-image generation and editing, outperforming state-of-the-art models in compositionality and semantic alignment.

Different Prompts and Their Outputs

Settings:

- For Image Generation: IterComp ⟨https://github.com/YangLing0818/IterComp.git⟩
- MLLM model used: DeepSeekR1



```
split_ratio=split_ratio, # The ratio of the regional prompt, the number of prompts is the same as the number of regions
batch_size = 1, #batch size
base_ratio = 0.5, # The ratio of the base prompt
base_prompt= prompt,
num_inference_steps=20, # sampling step
height = 1024,
negative_prompt=negative_prompt, # negative prompt
width = 1024,
seed = None,# random seed
guidance_scale = 7.0
```

Figure: Model configuration settings

Prompt Related to Medical Field

Prompt Given:MRI scan of a healthy human brain in T1-weighted contrast
Model Used: stable-diffusion-v1-4

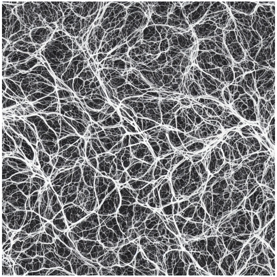


Figure: MRI scan of a healthy human brain in T1-weighted contrast

Prompt Related to Medical Field

Prompt Given: Showing the subtrochanteric fracture in the porotic bone.
Model Used: Nihirc/Prompt2MedImage

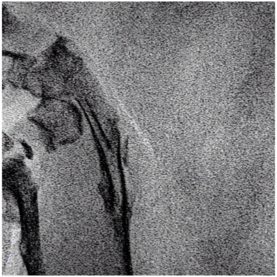


Figure: Bone Fracture

Prompt Related to Real Life Scenario

Prompt Given: ”A nighttime highway accident scene involving a luxury white Mercedes-Benz SUV crashed into a road divider, with the front of the vehicle severely damaged and smoke rising from the engine. The SUV is on fire, flames visible from the hood, and debris scattered across the road. A man in a torn sports tracksuit lies injured beside the vehicle, with visible bruises and blood on his forehead and arms. Passersby and a couple of concerned bystanders are seen trying to help. The surroundings include a dark rural highway, a broken fence, and a blurry ambulance approaching in the background.”

Model Used: IterCOMP

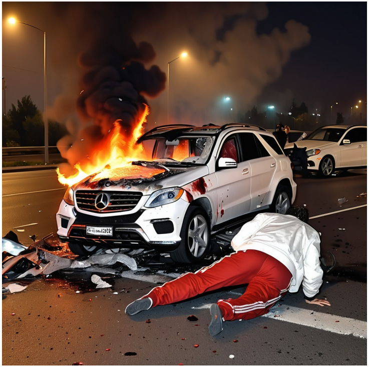


Figure: A Car Accident

Prompt Related Astronomy and Space

Prompt Given: A visualization of the merger of two neutron stars, 130 million light-years from Earth, based on data from the GW170817 gravitational wave event. The moment captures tidal distortions, spiral mass ejection, and jet formation leading to a short gamma-ray burst. In the background, a kilonova explodes, dispersing r-process elements such as strontium, gold, and europium into the interstellar medium. Scientists in the foreground analyze real-time LIGO and Virgo interferometry data visualized through dynamic spacetime curvature plots and particle tracking simulations.

Model Used: IterCOMP

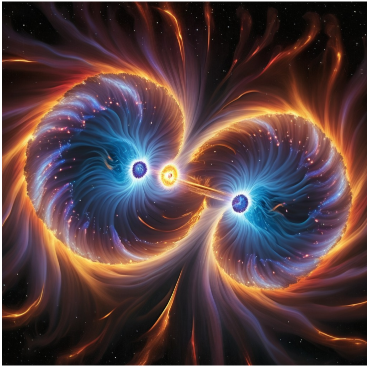


Figure: An Astronomy Example

Some Other Results

Figure: A Picnic Tour In Goa

Figure: Cat with Flowers

Motivation for Introducing Task-Specific Pipelines in RPG

From the above analysis, we infer that a **single model cannot handle all types of domain-specific prompts efficiently**, especially those involving complex object compositions.

We propose a pipeline that first performs **text classification**, and based on the classification, invokes **task-specific models** within the RPG (Recaption, Plan, Generate) framework.

Multimodal LLMs (MLLMs) show strong reasoning abilities, allowing decomposition of prompts into sub-tasks and enabling better performance via **chain-of-thought (CoT) planning**.

Motivation for Introducing Task-Specific Pipelines in RPG

The RPG framework supports **complementary regional diffusion**, where sub-regions are generated independently but harmoniously, tailored to individual subprompts.
It provides a **training-free** architecture with broad compatibility across various diffusion models (e.g., SDXL, DALL·E 3) and LLM backbones.

Handles **multi-category object composition and fine-grained spatial layout control** significantly better than single diffusion models.

Achieves **superior text-image semantic alignment** by iterating over recaptioning and layout planning in a loop, improving visual fidelity and alignment with prompt intent.

Thank You