Diffusion Detective: Revealing AI's Creative Process

Team Member:
Ankit Sanjyal · s
Course:
Deep Learning
Instructor:
Dr. Wenqi Wei

"Decoding AI's thinking, without the trench coat."

Motivation and Objectives

Large diffusion models can generate amazing images from text prompts, yet the reasoning and planning behind their creative process remains unknown. I ask myself, why does a model interpret "a cat wearing a crown" as majestic in one run and messy in another? This lack of transparency creates both trust issues (I can't predict outcomes) and slows down research (how do we debug or guide the model's reasoning).

This project aims to turn diffusion models from black boxes into detectives that narrate their own reasoning by proposing a lightweight, interpretable pipeline that fuses Chain-of-Thought (CoT) reasoning from language models with diffusion-based generation, enabling the system to explain what it will generate and why.

Key Questions asked:

- Can we make diffusion models describe their internal generative reasoning in natural language?
- Can we correlate reasoning traces (textual steps) with visual features (latent activations) for interpretability?
- Does reasoning supervision improve the controllability and stability of generation?

Related Work

Diffusion Models

I've always been fascinated by how Stable Diffusion models like sd1.5, sdxl and sd2.1 can turn a simple text prompt into a detailed image. Regardless of its impressive outputs, the process is a black box, I mean each latent step encodes important visual cues, but we can't see why the model made certain choices. This lack of interpretability sparked my curiosity: can we uncover and control the reasoning inside these models?

Chain-of-Thought and Vision-Language Reasoning

Recent works show that reasoning can be structured and interpretable. CoT-VLA(2025) breaks visual tasks into subgoals, UniFusion(2025) aligns text and vision effectively and MMada(2025) integrates reasoning across modalities. While these improve interpretability, they don't directly guide image generation. This project extends these ideas to SD2.1, letting reasoning traces actively influence the generated images.

Explainability and Control

Building on models like TuneFree-CFG (2025) {My own Paper under review} and Diffusion Policy (2024), Diffusion Detective introduces a Reasoning-Aware Guidance DScheduler that adjusts how reasoning steps influence the denoising process. Users should be able to see which reasoning step affected each image region and even edit steps interactively, transforming a black-box model into a controllable, interpretable creative tool.

Proposed Work

System Overview:

The Diffusion Detective integrates three key components:

1) Reasoning Generator (language module):

   Uses CoT-VLA or LLAVA-Next to decompose the text prompt into a structured reasoning chain. For example [Perhaps]:

   Prompt: "A detective examining clues in a foggy alley."

   Reasoning Chain: [Detective -> Coat -> Streetlight -> Fog -> Suspense Tone]

2) Reasoning to visual grounding (fusion module):

   Adapts Unifusion's multimodal alignment layer to bind each reasoning step to a diffusion latent representation and each step produces a partial latent "evidence map" representing incremental visual understanding.

3) Iterative Diffusion with reasoning-aware guidance (generation module):

   Uses Stable Diffusion 2.1 as the visual backbone and introduces a Reasoning-Aware CFG scheduler that dynamically adjusts textual guidance weights based on the semantic complexity of reasoning steps. This enables interpretability i.e, users can see how the model's reasoning affects each visual layer.

4) Interactive Demo: "Diffusion Detective Studio per say":

   The gradio built interface should allow users to:

   - Input Prompts
   - Observe generated reasoning steps
   - Visualize the image evolving step-by-step
   - Edit reasoning steps mid-process and see updated results in real time

System Architecture:

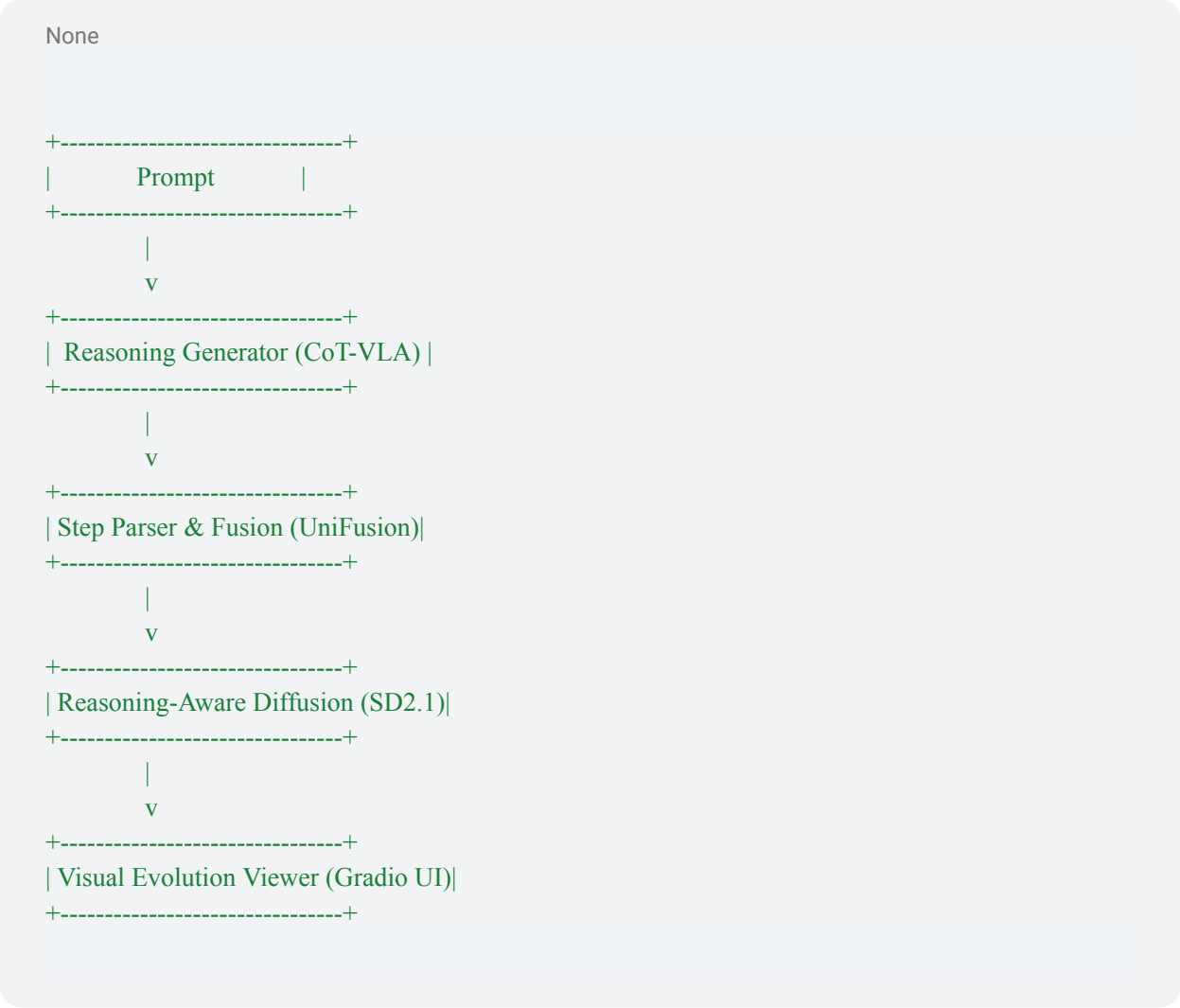

Fig: System Architecture Drawn by Gemini2.5

Novelty

Unlike these large-scale systems, this project focuses on clear, step-by-step reasoning within a smaller, easy to run setup that works well for class projects but can also be expanded for full research.

Plan Of Action

I plan to incrementally build the project over the next month. I will start by integrating a chain-of-thought reasoning model to decompose input prompts into structured reasoning steps. Those steps will then be aligned with the diffusion latent space using a multimodal fusion module inspired by UniFusion, creating intermediate visual evidence for each reasoning stage. Then I will implement the Reasoning-Aware Guidance Scheduler within stable diffusion 2.1 to control how each reasoning step influences the iterative

denoising process. Then finally, I will build an interactive Gradio demo that visualizes the reasoning chain and stepwise image evolution, allowing users to intervene and modify reasoning steps in real time. At first it sounds overwhelming, but with the availability of vibe coding, it should be doable.

Resources:

- Models: Stable Diffusion v2.1, Mistrail-7B-Instruct, CLIP/SigLIP
- Data: DiffusionDB(Prompt-image pairs), COCO Captions (alignment fine-tuning), synthetic reasoning traces from GPT
- Hardware: Colab Pro/ MPS/ HPC
- Frameworks: Pytorch, Diffusers, Gradio, Transformers, HuggingFace

Weekly Schedule:

Week	Task	Deliverable
1	Environment setup, Stable Diffusion integration	Working pipeline generating images from prompts
1	Extract latent activations and attention maps	Visualizations of intermediate features
2	Generate textual reasoning trace per diffusion step	Notebook outputting text + attention heatmaps
2	Build interactive interface (Gradio/Streamlit)	GUI showing prompt → reasoning → image
3	Implement human-guided edits (object positions, emphasis, style)	Interactive demo where edits modify outputs
3	Evaluate results: CLIP similarity, MS-SSIM, human study	Plots/tables comparing vanilla vs guided diffusion
4	Ablation studies, finalize figures	Clear visualizations for class presentation
4	Prepare final project report and demo	3–5 page report + working demo ready for showcase

Evaluation and Testing Method

The evaluation will run in two axes:

1) Interpretability Metrics: Clip similarity to measure coherence between reasoning trace and final image and attention trace correlation to measure how strongly each reasoning step aligns with specific U-Net activations
2) Generation Quality: FID and CLIP-IQA to measure image quality and semantic similarity,Human Study if time permits to explain how reasonable the model's explanation sounds and a demo validation to measure success when altering reasoning text predictably changes image output.

Bibliography

[1] Zhao et al., *"CoT-VLA: Chain-of-Thought Vision-Language-Action Models,"* arXiv:2503.22020, 2025.

[2] Li et al., *"UniFusion: Unified Multimodal Fusion for Generalizable Reasoning,"* arXiv:2510.12789, 2025.

[3] Yang et al., *"MMada: Multimodal Adaptive Alignment for Reasoning-Driven Tasks,"* arXiv:2505.15809, 2025.

[4] Rombach et al., *"High-Resolution Image Synthesis with Latent Diffusion Models,"* CVPR 2022.

[5] Hertz et al., *"Prompt-to-Prompt Image Editing with Cross-Attention Control,"* SIGGRAPH 2023.

[6] Radford et al., *"Learning Transferable Visual Models from Natural Language Supervision,"* ICML 2021.