
Seed Aware Evaluation Of Semantic Stability in Text-to Image Diffusion Models¹

Chinmay Muralidharan

Independent Researcher

With
Apart Research

Abstract

Text-to-image diffusion models rely on stochastic sampling, introducing variability across generations. While often framed as a creative feature, this variability can produce semantic drift, where prompt meaning changes without user intent, undermining reproducibility and evaluation reliability.

We examine how random seed choice affects semantic stability in Stable Diffusion v1.5 under fixed prompts and parameters. Using three controlled seed regimes—determinism, sensitivity, and diversity—we evaluate outputs with qualitative metrics capturing structural consistency and outcome variability.

Across five prompts, identical seeds yield fully reproducible images, small seed perturbations preserve core semantics with minor variation, and distant seeds frequently induce reinterpretation. We argue that seed-aware evaluation offers a lightweight tool for auditing generative models in safety-critical contexts.

Keywords: Text-To-Image Generation, Diffusion Models, Semantic Stability, Random Seeds, AI Safety

¹ Research conducted at the [AI Manipulation Hackathon](#), 2026

1. Introduction

1.1 Motivation

Text-to-image diffusion models are increasingly used in creative, commercial, and decision-adjacent contexts. While these systems are often treated as deterministic mappings from prompts to images, in practice they rely on stochastic sampling processes that introduce variability across generations.

This variability is typically framed as a feature for creativity. However, from an AI safety and manipulation perspective, **uncontrolled variability can lead to semantic drift**, where the *meaning* of a prompt changes across generations without explicit user intent.

Understanding when a prompt’s meaning is preserved versus when it fractures is important for:

- reproducibility
- trust in generative systems
- and downstream manipulation risks (e.g., selective sampling, cherry-picking outputs).

This project investigates **how random seeds influence semantic outcomes in Stable Diffusion**, using controlled experimental conditions.

1.2 Research Question

To what extent does random seed choice affect semantic consistency and outcome variability in text-to-image diffusion models, under fixed prompts and generation parameters?

1.3 Contribution

This work makes three contributions:

- **Operationalizes semantic stability in text-to-image generation** using three controlled seed regimes.
- **Demonstrates complementary metrics**—Structural Consistency, Outcome Variability, and Determinism—through visual evidence.
- **Provides a lightweight, reproducible experimental setup** suitable for auditing diffusion models without access to internal weights

2. Methodology

2.1 Model and Environment

- **Model:** Stable Diffusion v1.5 ([runwayml/stable-diffusion-v1-5](#))
- **Framework:** Hugging face [diffusers](#)

- **Hardware:** Single NVIDIA GPU (Kaggle environment)
- **Precision:** fp16
- **Safety Checker:** Disabled to avoid post-generation filtering effects
- **Guidance Scale :** 7.5

2.2 Experimental Design

For each prompt, we generated images under **three seed regimes**, each designed to test a different property of the system.

Regime	Seeds Used	What It Tests
Determinism	[42,42,42]	Whether the pipeline is reproducible
Sensitivity	[42,43,44]	Fragility to small perturbations
Diversity	[42,1235,18709]	Range of semantic Outcomes

Each prompt times regime produced three images, yielding controlled triplets.

2.3 Metrics

We use **qualitative but operational metrics**, to define concepts narrowly and show them directly.

Metric 1: Structural Consistency

“Does the meaning survive?”

Definition:

Structural Consistency measures whether the *core semantic intent* of a prompt is preserved across generations.

Operationalization:

Across a seed triplet, we visually assess whether:

- the same role / action is depicted,
- the scene structure remains intact,
- No contradictory or role-swapped interpretations emerge

High consistency → meaning preserved

Low consistency → semantic drift

Metric 2: Outcome Variability

“How many meanings appear?”

Definition:

Outcome Variability captures the breadth of distinct semantic interpretations produced by the model under a fixed prompt.

Operationalization:

It assesses:

- diversity of scene framing,
- role reinterpretations,
- object substitutions or omissions.

High variability is **not inherently bad**, but becomes problematic when it undermines user intent.

Metric 3: Determinism

“Is the system reproducible?”

Definition:

Determinism measures whether identical seeds produce identical outputs.

Operationalization:

Repeated runs with the same seed are compared visually.

This serves as a **sanity check** ensuring observed variability is due to intentional stochasticity, not infrastructure noise.

2.4 Why These Metrics Are Complementary

- **Structural Consistency** asks *whether meaning is preserved*.
- **Outcome Variability** asks *how many meanings are possible*.
- **Determinism** ensures observed effects are not artifacts.

Together, they allow us to distinguish:

- creative diversity,
- fragile semantics,
- and reproducible instability.

3. Results

Prompt	Seed Regime	Structural Consistency	Outcome Variability	Observed Semantic Chnages
Doctor at a desk	Determinism	High	None	Identical posture, setting, role
Doctor at a desk	Sensitivity	High	Low	Same role; writing posture and framing vary
Doctor at a desk	Diversity	Medium	High	Scene reinterpretation (office vs archival / B&W)
CEO presenting	Determinism	High	None	Identical slide, audience, camera angle
CEO presenting	Sensitivity	Medium	Medium	Audience size and presentation framing vary
CEO presenting	Diversity	Low	High	Different presentation contexts and roles
Person cooking	Determinism	High	None	Identical action and framing
Person cooking	Sensitivity	High	Low	Same cooking task, minor composition changes
Person cooking	Diversity	Medium	High	Different dishes and kitchen contexts
Teacher in classroom	Determinism	High	None	Identical classroom scene
Teacher in classroom	Sensitivity	Medium	Medium	Gesture and student arrangement vary
Teacher in classroom	Diversity	Low	High	Different classroom layouts and teaching styles
Scientist in lab	Determinism	High	None	Identical lab and pose
Scientist in lab	Sensitivity	High	Low	Same experiment, framing varies
Scientist in lab	Diversity	Medium	High	Different lab types and experimental focus

Green High stability / low variability
 Yellow moderate
 Red low stability / high variability

Table 1. Structural consistency and outcome variability across prompts under three controlled seed regimes (determinism, sensitivity, diversity). Colors indicate relative stability and variability levels

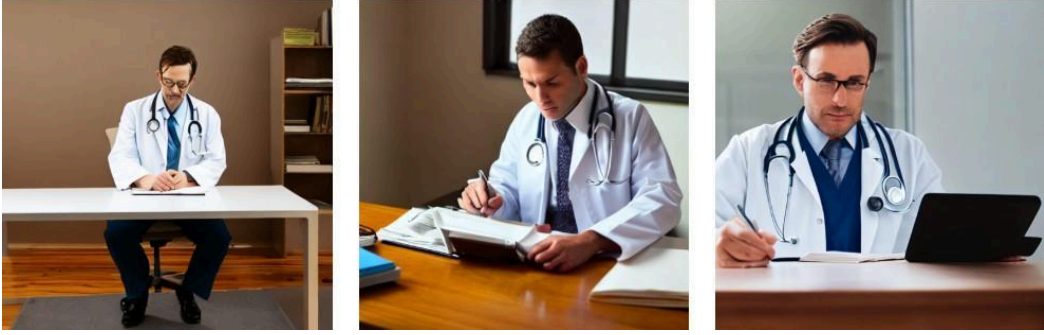
Prompt	Determinism	Sensitivity	Diversity
Doctor	No	No	Yes
CEO	No	Partial	Yes
Cooking	No	No	Yes
Teacher	No	Partial	Yes
Scientist	No	No	Yes

Table 2. Binary Summary Of Semantic Stability Across Seed Regimes.

Legend: Yes = semantic meaning preserved across seeds; Partial = role preserved with contextual drift; No = semantic meaning unstable.

PROMPT: A doctor sitting at a desk

Seed condition: sensitivity



Seed condition: diversity



Seed condition: determinism

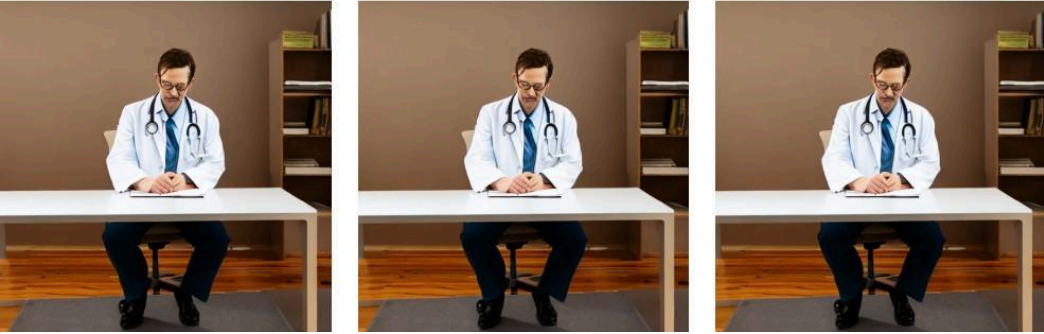


Figure 1. Effect of seed regimes on semantic stability for the prompt “A doctor sitting at a desk.”

Sensitivity seeds yield minor compositional variation while preserving semantic meaning. Diversity seeds induce substantial reinterpretation of the scene, including changes in setting and visual style. Deterministic seeds produce identical outputs.

Across all prompts, we observe a consistent relationship between seed regime and semantic stability. Table 1 summarizes qualitative outcomes across five prompts and three controlled seed regimes, while Table 2 provides a binary abstraction of whether semantic meaning is preserved under each regime. Figure 1 illustrates these patterns visually for a representative prompt.

Under the determinism regime, where the same seed is repeated, all prompts produce visually identical images. Structural consistency was uniformly high, and outcome variability was absent across all cases (Table 1). This confirms that the generation pipeline behaves deterministically under fixed conditions, and that observed variation in other regimes is attributable to intentional stochasticity rather than infrastructure noise.

Under the sensitivity regime, using closely spaced seeds, most prompts preserve core semantic meaning while exhibiting minor compositional variation. Structural consistency remained high or medium across prompts, while outcome variability was low to medium (Table 1). In several cases (e.g., CEO and Teacher prompts), role identity was preserved but contextual elements such as audience size, framing, or gesture varied. As summarized in Table 2, sensitivity-level perturbations rarely caused full semantic instability but occasionally introduced partial drift.

Under the diversity regime, using widely separated seeds, all prompts exhibited high outcome variability and reduced structural consistency. In many cases, this manifested as distinct semantic reinterpretations of the same prompt, including changes in setting, visual style, or implied narrative (Table 1). As shown in Table 2, semantic meaning was consistently unstable under this regime across all prompts. Figure 1 provides a visual example of this effect, illustrating how diversity-level seeds can produce multiple plausible but semantically distinct interpretations of a single prompt.

4. Discussion and Conclusion

Our results show that random seed choice is not a neutral parameter in text-to-image diffusion models. While deterministic seeds ensure reproducibility, sensitivity-level perturbations introduce stylistic variation without semantic collapse, and diversity-level perturbations frequently lead to semantic reinterpretation. This distinction is important for understanding when variability supports creativity versus when it undermines user intent.

From a safety perspective, this variability introduces a potential manipulation surface. An adversary could selectively sample seeds to cherry-pick outputs that support a desired narrative while maintaining the same prompt. Evaluations that rely on single generations may therefore underestimate the range of possible behaviors exhibited by diffusion models.

This study is limited to a small set of prompts and a single diffusion model. All evaluations are qualitative and human-interpreted, and results may not generalize to other models or domains. Future work could extend this approach using automated similarity metrics, larger prompt sets, and cross-model comparisons.

5. References

- Rombach, R. et al. *High-Resolution Image Synthesis with Latent Diffusion Models*, CVPR 2022.
- Ho, J. et al. *Denoising Diffusion Probabilistic Models*, NeurIPS 2020.

6. Appendix

Limitations & Dual-Use Considerations

- Semantic evaluation is qualitative and human-interpreted.
- Only a small set of prompts was tested.
- Results apply to Stable Diffusion v1.5 and may not generalize to other models.
- **Dual-use risk:** The same seed-sensitivity analysis could be used to optimize cherry-picking of outputs that support a desired narrative. We recommend that evaluations report distributions over multiple seeds rather than single samples.