Probing Physical Realism in Diffusion Image Generators

Stanford CS229 Project

Sarang Goel **Arnav Chakravarthy** **Ekansh Mittal**
Department of Computer Science, Stanford University
sarang28@stanford.edu arnavak@stanford.edu ekanshm@stanford.edu

1 Introduction

Modern image generation models can produce hyper-realistic scenes akin to real-world phenomena; however, they also excel at producing visually harmonious scenes that are impossible to produce in the real world, such as floating unsupported objects, inconsistent object reflections, and off-center balanced objects. This project aims to determine whether a modern diffusion model possesses internal representations that can distinguish between image generations that adhere to everyday physics and those that violate it, beyond superficial signals such as text rarity or low image quality. Understanding this internal distinction would deepen our understanding of what these models learn about the world and whether they can grasp abstract concepts such as realism, benefiting tools for safety in generative systems.

To investigate this, we decided to focus on the inner workings of flow-diffusion models, particularly the FLUX.1-schnell model. Schnell is a heavily distilled model, meaning that it is capable of generating high-fidelity images with just four inference steps compared to the usual 20-30 inference steps most other diffusion-based image generation models employ. Due to the short inference stage, the model is forced to align with physical constraints quickly and aggressively. We purposely chose this heavily distilled model architecture as the industry has increasingly been moving towards low-latency, real-time applications for image generation, such as edge computing, interactive media, and robotic simulations, where models can not rely on long refinement loops to generate images. As these fast generative models become more mainstream, the ability to discern physical hallucinations from generated images is an imperative safety and reliability concern.

To evaluate the model's physical understanding, we generated a dataset of 1,020 images across three physical violation categories we defined: gravity, reflection, and balance. We then trained linear probes on the model's internal activations to further classify the physically plausible versus implausible generations.

2 Related Work

Our work builds on two approaches within generative AI: evaluating physical realism and probing internal representations. In the past, physics evaluation required human triage, which was later replaced by methods that utilized automation. For example, Yuan et al. (2026) uses the denoising likelihood to evaluate video models, while Banani et al. (2024) evaluates zero-shot 3D spatial consistency. However, these methods are limited by the fact that they treat generative models as a "black box", and only focus on the final outputs and rely on synthetic data simulators. Other works have attempted to open the model and directly apply linear probing to the model's activations. These works include Chen et al. (2023), which utilizes linear probing to identify 3D scene depth in Stable Diffusion, as well as Sclocchi et al. (2024) and Al-Ramahi et al. (2024) that attempt to track gradual phase transitions over long generations (50-100 steps). While the strength of these works lie in their capability of localizing abstract concepts, their weakness lies in their inability to disentangle actual spatial reasoning from the text conditioning. We believe our approach addresses this as we have designed a text-only baseline to account for any semantic leakage, and have performed a manual

triaging process to ensure that there exists high visual fidelity in the inputs used. Unlike previous work that have focused on classical Latent Diffusion Models, our work also explores representational probing with FLUX.1-schnell to determine if geometric representations also exists in low-latency, real-time generative models.

3 Dataset and Features

Since we needed to probe the internal representations of the generative model, we created a custom dataset of images with the FLUX.1-schnell model. This dataset consisted of 1,020 images (samples shown in Figure 1 with a resolution of 1024x1024 pixels and across three physical properties: gravity, reflection, and balance. To generate these image, we used a prompt matrix (shown in Table 1) consisting of 17 subjects and 4 relations per category (2 plausible and 2 implausible physical constraints), totaling 204 unique text prompts.

(a) Gravity (Plausible)

(b) Reflection (Plausible)

(c) Balance (Plausible)

(d) Gravity (Implausible)

(e) Reflection (Implausible)

(f) Balance (Implausible)

Figure 1: Examples from our generated FLUX.1-schnell dataset demonstrating minimal pairs across the three physical categories. The top row shows physically plausible generations (Label: 0), while the bottom row shows physically implausible generations (Label: 1).

Then, we passed these prompts through the model using 5 different random seeds (42, 100, 404, 888, and 1234) to ensure diversity in the latent representations, giving 340 image generations per physical category.

Table 1: Combinatorial prompt matrix for dataset generation. Each prompt is constructed using the template: "[Article] [Subject] [Relation]". Minimal pairs ensure that semantic structure remains identical while isolating the physical violation.

Violation Type & Example Subjects	**Plausible Relation (Label: 0)**	**Implausible Relation (Label: 1)**
Gravity (mug, apple, cat, book, chair. . .)	resting on a table	floating above a table
Reflection (mug, apple, cat, book, chair. . .)	placed off-center before a mirror, showing its normal reflection clearly in the glass	placed off-center before a mirror, but magically showing absolutely no reflection in the glass
Balance (mug, apple, cat, book, chair. . .)	resting flat and stably in the center of a wooden table	balancing impossibly on its very edge on the sharp tip of a sewing needle

Since the model sometimes failed to render the physical structure explained by the prompt, we also underwent a manual triaging protocol, where any generation that had distorted visuals or missed the specified constraints were removed from the dataset. This ensured that the linear probes evaluated

physical plausibility rather than low quality generations, which are can occur when prompts describe physically implausible scenarios.

Rather than raw pixels, our algorithm inputs consist of pre-computed internal neural activations. For our image features, we attached a forward hook to the 10th transformer block (single_transformer_blocks[10]) of the 4-step FLUX.1-schnell model to capture the intermediate hidden states across all four inference timesteps. We then applied global average pooling across the spatial dimension, compressing the multidimensional tensors into computationally viable 1-dimensional embeddings of size $\mathbb{R}^{3072}$. As a baseline to control for semantic leakage, we also extracted frozen text embeddings from the model's text encoders. Finally, our data splitting varied by experiment: within-category and global tests, we used a 80/20 train/test split alongside a 5-fold cross-validation, whereas for the holdout generalization tests we used a categorical split by training on two physical categories and evaluating on the unseen third.

4 Methods

4.1 Fundamentals: Logistic Regression

We used logistic regression to evaluate the linear seperability of physical plausibility for both text and image representations. Logistic regression models the probability that a given input x belongs to the positive class ($y = 1$). We define the positive class to be "implausible" and the negative ($y = 0$) class to be "plausible.

In order to produce a probability, logistic regression passes a linear combination of the input features through the sigmoid function, which maps the output to the range $(0, 1)$. Sigmoid for a model parameratized by weights θ is defined as $g(z) = \frac{1}{1+e^{-z}}$. Then, the sigmoid for a model parameterized by weights θ is defined as $h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$.

We minimized the regularized binary cross-entropy cost function in order to find θ. In order prevent overfitting, we apply L2 regularization (penalty parameter λ). Since we have a very low sample count relative to the dimensionality of our embedding space, this is critical. The objective function is defined as $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right] + \frac{\lambda}{2}\|\theta\|^2$.

By minimizing this objective function logistic regresion finds the best hyperplane in feature space to serve as a decision boundary between physically plausible and implausible representations.

4.2 Text-Embedding Baseline

Before analyzing the diffusion model's internal embeddings, we established a text only baseline to determine how seperable the linguistic cues of plausibility and implausibility are. To do this we extracted the 200 unique prompt templates from our dataset, and process them through a pre-trained SentenceTransformer (all-MiniLM-L6-v2). This represents the text prompt as a vector in a 384-dimensional vector space. As described previously, we define the target variable to be the plausibility of the prompt.

We then train a logistic regression classifier on these text embeddings to predict the plausibility of the prompt. To do this, we partitioned the 200 encoded samples into an 80% training set and 20% training set, using a stratified split to ensure that a 50/50 class distribution is maintained in both sets. Since our deduplicated sample size is small ($N = 200$), a single train-test split would result in a high-variance estimate of performance. To ensure that our evaluation of the linear seperability is robust, we implemented a 5-fold stratified cross-validation. We implemented this by splitting the datset into 5 folds of 40 samples each (maintaining a uniform class distribution in each). For each of the 5 folds, the logistic regression model is trained on the remaining four folds an evaluated.

4.3 Linear Probing on Diffusion Activations

To test our hypothesis, we train logistic regression probes on the pooled internal hidden states extracted from the diffusion model across multiple denoising timesteps.

To ensure that the probes are not learning class imbalances we ensure that for each specific model, the plausibility classes are balanced. For all probes, we randomly down sample the majority class

to match the minority class. Similar to the text baselines, we trained all linear probes using 5-fold stratified cross-validation to predict the plausibility of the embeddings. Prior to training, the image embeddings are standardized to a mean of zero and variance of one using a scaler. We trained three sets of L2-regularized logistic regression probes. The first was trained on unique timestep/violation type pairs (12 total) to understand when specific physical constraints are resolved. The second set was trained on data combined across the full dataset violation types for each timestep (4 total) to understand if the probes can find an abstraction across them. The third set was trained on two violation categories and tested on the third, unseen violation category, serving as a holdout generalization test to determine if the the classifiers are able to learn a universal representation of physical realism.

5 Experiments / Results / Discussion

5.1 Evaluation Metrics

We report two primary metrics: **AUROC** and **macro-averaged F1**.

The ROC curve plots the True Positive Rate against the False Positive Rate as the classifier's decision threshold is swept from 0 to 1, where $TPR = TP/(TP+FN)$ and $FPR = FP/(FP+TN)$. The AUROC is the integral of this curve, equal to the probability that the model ranks a randomly chosen positive (implausible) example above a randomly chosen negative (plausible) one: $AUROC = \Pr[\hat{f}(x^+) > \hat{f}(x^-)]$. A score of 1.0 is perfect; 0.5 is random guessing. We prefer AUROC over accuracy because our held-out splits are class-imbalanced (reflection: 42 implausible vs. 116 plausible). Macro-averaged F1 averages the per-class score $F_k = 2P_kR_k/(P_k + R_k)$ equally across classes, guarding against classifiers that exploit the majority class.

5.2 In-Distribution and Combined Probing

Linear probes were trained within each violation and achieved perfect or near-perfect separation for all timesteps (AUROC $= 1.000$ for gravity and balance, reflection degrades slightly to 0.983 at step 3). A combined probe trained on a all the violation types achieved an AUROC $= 1.000$ at every timestep as well, with only one gravity misclassification at step 0 ($F_1 = 0.980$). This suggests a shared latent "plausibility direction" across differing physical representations.

5.3 Zero-Shot Generalization to Unseen Violation Types

We trained on two violation types and tested on the held-out third at all four denoising timesteps with both logistic regression and linear SVM. Results are shown in Figure 2. Reflection generalizes best (logistic regression AUROC $= 0.911$ at step 2); balance peaks at 0.955 (step 0) but decays to 0.753 by step 3; gravity is the hardest held-out target, dropping from 0.894 to 0.629 for logistic regression and near the random chance threshold for the SVM at step 3. Early timesteps generalize better than later ones across all settings. Figure 3 visualizes the increasing distribution overlap in held-out conditions, mirroring the declining AUROCs.

5.4 Text Semantics Baseline

To control for the possibility that probes exploit text conditioning rather than visual geometry, we trained an identical logistic regression on 384-dimensional `all-MiniLM-L6-v2` sentence embeddings of the prompts alone. In-distribution, the text probe achieves 100% test accuracy (5-fold CV: $99.02\% \pm 1.19\%$). Under the same leave-one-out zero-shot protocol, the text baseline scores AUROC $= 0.970$ (gravity), 1.000 (balance), and 0.835 (reflection), compared against visual activations in Table 2 and Figure 4.

Table 2: Zero-shot generalization AUROC: text embeddings vs. visual activations (logistic regression, Step 2).

Method	Gravity	Balance	Reflection
Text Baseline (MiniLM)	0.970	1.000	0.835
Visual Activations (Step 2)	0.704	0.800	0.911

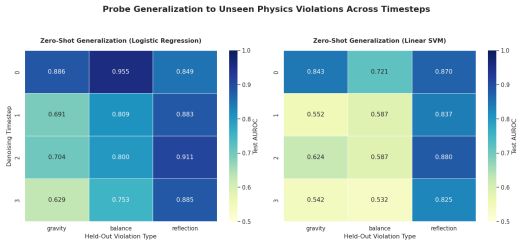

Figure 2: Zero-shot generalization AUROC across held-out violation types and timesteps for logistic regression (left) and linear SVM (right).

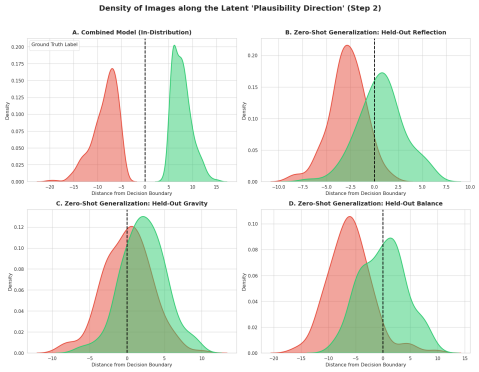

Figure 3: Latent plausibility direction at Step 2: (A) in-distribution, (B) held-out reflection, (C) held-out gravity, (D) held-out balance. Red = implausible, green = plausible.

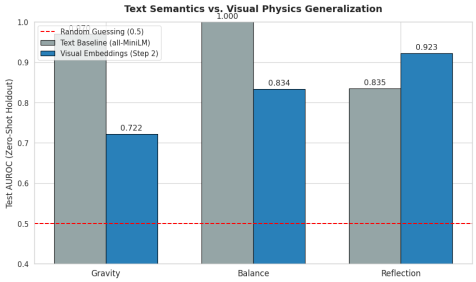

Figure 4: Zero-shot AUROC: text baseline (gray) vs. visual embeddings at Step 2 (blue). Dashed red = random guessing (0.5).

The text baseline leads on gravity and balance, where explicit verbal markers (“floating,” “off-center”) provide strong semantic signal. However, for held-out reflection, visual activations (0.911) substantially outperform the text baseline (0.835), indicating the model’s internal representations encode geometric information about mirror consistency that cannot be decoded from language alone. Both methods exceed random guessing on all three held-out types, confirming the visual plausibility signal is not merely a residual of text conditioning.

6 Conclusion / Future Work

From our results, we found that data within violation types is perfectly separable for both text and image embeddings. To further evaluate physical understanding, we performed zero-shot generalization experiments. For categories like gravity and balance, the text baseline outperformed the visual probes, which may be because these violations rely strong on verbal cues like "floating", which provide strong signals in the language space itself. However, for held-out reflection violations, the visual probe achieved a high AUROC of 0.911, which outperformed the baseline of 0.835. This indicates that the diffusion model encoded some structural representation of reflection consistency that could not be decoded from language alone. We additionally found that embeddings from early denoising steps were more linear separable, which suggests that abstract physical concepts are resolved earlier in the denoising process, before the later denoising steps start to get dominated by finer visual details.

If given more time and resources, our main future direction would be to perform interventional experiments. We would like to see if nudging intermediate activations across the learned vector of plausibility from the probes can correct the model’s physical violations during the denoising process. We would also like to expand our dataset to include more complex physical phenomena, like shadows and fluid dynamics.

7 Contributions

- **Arnav Chakravarthy**: Configured FLUX.1-schnell generation pipeline with memory offloading in Drive, executed programmatic generation of the 1000 image dataset, drafted hook methodology for upcoming activation extraction, drafted Results sections
- **Sarang Goel**: Established cross-validated text-embedding baseline, ensured proper data-leakage prevention across generated seeds, and drafted Introduction section
- **Ekansh Mittal**: Implemented linear probing experiments, designed combinatorial prompt matrix for dataset, established/executed manual triage protocol ensuring data validity, and drafted Methods and Conclusions section

References

Mohammad Al-Ramahi, Izzat Alsmadi, and Abdullah Wahbeh. 2024. Predicting question quality on stackoverflow with neural networks.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. 2024. Probing the 3d awareness of visual foundation models.

Yida Chen, Fernanda Viégas, and Martin Wattenberg. 2023. Beyond surface statistics: Scene representations in a latent diffusion model.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. 2024. A phase transition in diffusion models reveals the hierarchical nature of data.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. 2026. Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference.