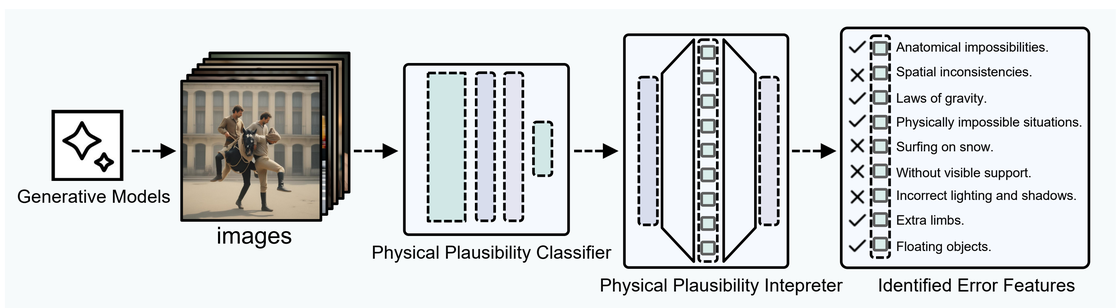
HOW DOES MY MODEL FAIL? AUTOMATIC IDENTIFICATION AND INTERPRETATION OF PHYSICAL PLAUSIBILITY FAILURE MODES WITH MATRYOSHKA TRANSCODERS

Yiming Tang,[1] Abhijeet Sinha[1] Dianbo Liu,[1]
[1] National University of Singapore

INTRODUCTION

Generative models produce photorealistic images that routinely violate basic physics — extra fingers, floating objects, malformed text. Aggregate metrics (FID, CLIPScore) and even SOTA LMMs miss these violations, and *no framework tells us which physical errors a model makes*. We introduce **Matryoshka Transcoders**, an interpretability method that automatically discovers and names physical-plausibility failure modes, then uses them to benchmark 8 SOTA generative models. Beyond benchmarking, the same pipeline generalizes to interpret any vision encoder.

FIGURE 1

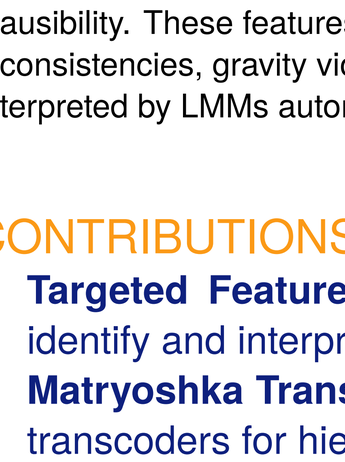


Matryoshka Transcoders for Automatic Physical Plausibility Analysis in Generative Models. We pass images from generative models to a physical plausibility classifier, whose representations are analyzed by Matryoshka Transcoders to extract hierarchical sparse features related to physical plausibility. These features encompasses anatomical impossibilities, spatial inconsistencies, gravity violations, and other physical implausibilities and are interpreted by LMMs automatically.

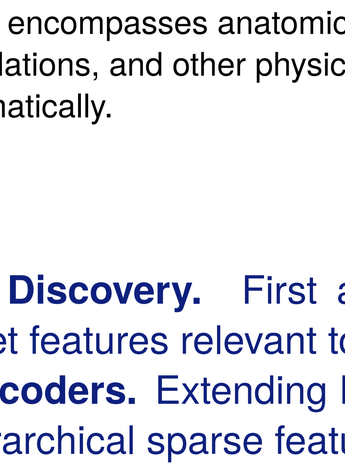
CONTRIBUTIONS

- **Targeted Feature Discovery.** First automated pipeline to identify and interpret features relevant to a specific topic.
- **Matryoshka Transcoders.** Extending Matryoshka training to transcoders for hierarchical sparse features.
- **Relevance Metrics.** Population- and description-based scores for topic alignment of discovered features.
- **Failure Insights.** Diverse, named violation modes revealing how each model breaks physics.

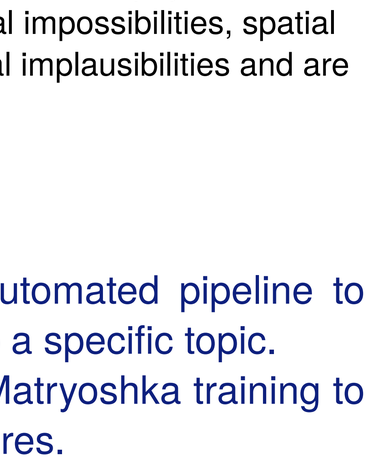
FIGURE 2



Feature 104
Theme: Tennis Related Activities.
Error Explanation: The figures in the images have distorted anatomies.
Error Type: Anatomy



Feature 243
Theme: The Existence of Textual Information.
Error Explanation: The texts in the images are distorted and not well-formed.
Error Type: Text-distortion



Feature 648
Theme: Handdrawing or Painting Styles.
Error Explanation: The images possess styles that identify them as implausible.
Error Type: Style

Qualitative examples of physical plausibility features discovered by Matryoshka Transcoders. Each column shows a representative feature with its top-activating images, automatically generated theme, error explanation, and categorized error type.

TABLE 2

(a) Feature relevance vs. baselines

Method	R_{pop}	R_{desc}
SAE	6.54%	6.01%
Matryoshka SAE	9.88%	5.66%
Transcoder	12.74%	11.52%
Matryoshka Transcoder	**16.00%**	**13.04%**

(b) Physical plausibility of 8 generative models

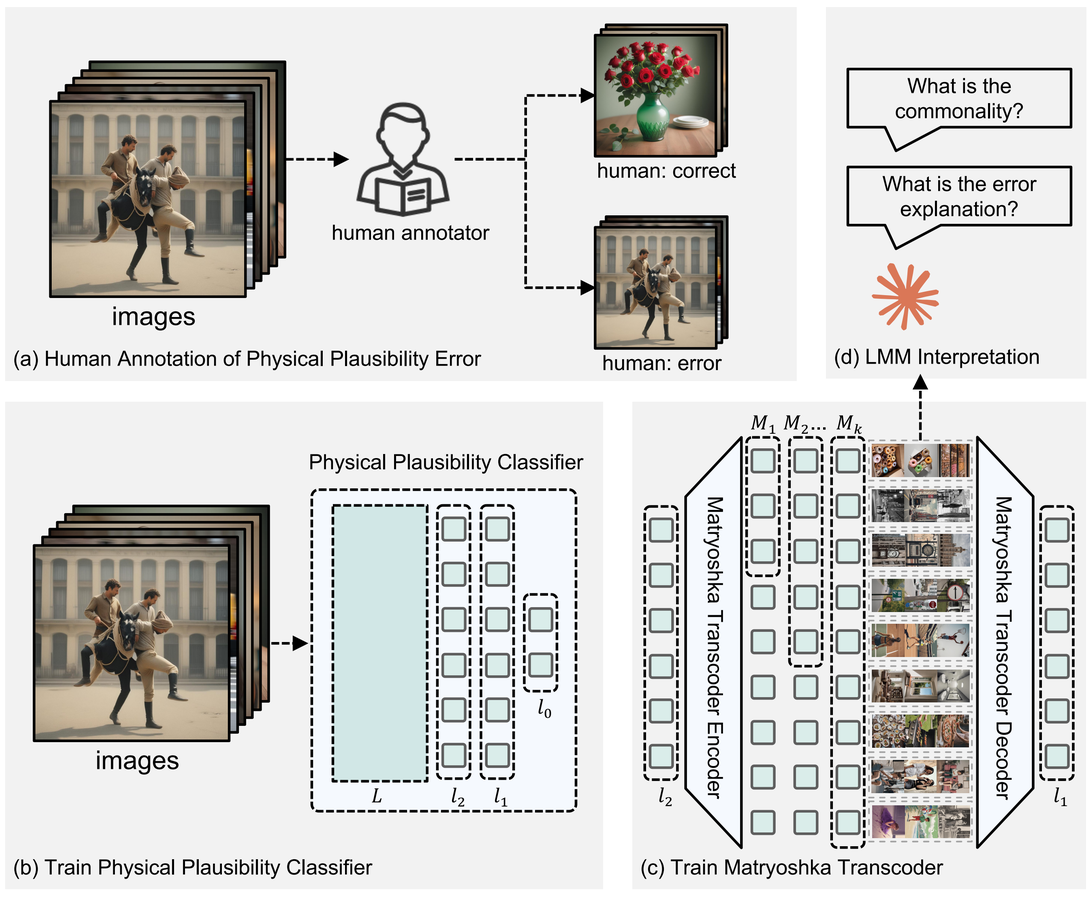
Model	Err ↓	Model	Err ↓
SDXL-turbo	74.2	SDXL-large	46.8
SDXL-base	53.1	Stable-Cascade	43.1
DALL·E 3	49.8	Kolors	40.8
SDXL-medium	48.7	**FLUX**	**38.0**

METHOD

We first build a labeled dataset of physical plausibility. Human annotators assign each generated image a binary label — *plausible* or *error* — guided by written instructions and representative examples of common violation types (anatomical impossibilities, distorted structures, gravity violations, malformed text). The plausible class is augmented with natural images from MSCOCO. On this dataset we train a binary classifier: a frozen CLIP-ViT-Large/14 encoder feeds a lightweight two-layer head ($768 \to 256 \to 1$), optimized with AdamW. Freezing CLIP preserves its rich vision–language representations while the head reads out the plausibility signal.

Matryoshka Transcoders then decompose the classifier's hidden activations into a sparse, interpretable dictionary of features. An encoder maps each activation into a high-dimensional latent space; a top-k constraint keeps only a few features active; a decoder reconstructs the activation. The Matryoshka twist: we train one shared dictionary under a *nested* sequence of sparsity budgets $\mathcal{M} = \{128, 256, 512, 1024, 2048\}$, demanding accurate reconstruction at every m simultaneously. Small budgets capture the most concentrated, reliable features; large budgets cover rare error modes — in a single model.

FIGURE 3



Complete pipeline for Matryoshka Transcoders. Our approach combines four stages: (a) human annotation of correct/error images, (b) training a physical plausibility classifier that possess relevant information, (c) training Matryoshka transcoders to discover sparse, interpretable features at nested granularities from classifier activations, and (d) using large multimodal models to interpret discovered features by identifying visual commonalities and error explanations from maximally activating images.