How does My Model Fail? Automatic Identification and Interpretation of Physical Plausibility Failure Modes with Matryoshka Transcoders

Yiming Tang
National University of Singapore
yiming@nus.edu.sg

Abhijeet Sinha
National University of Singapore
abhijeet@nus.edu.sg

Dianbo Liu
National University of Singapore
dianbo@nus.edu.sg

Abstract

Although recent generative models are remarkably capable of producing instruction-following and realistic outputs, they remain prone to notable physical plausibility failures. Though critical in applications, these physical plausibility errors often escape detection by existing evaluation methods. Furthermore, no framework exists for automatically identifying and interpreting specific physical error patterns in natural language, preventing targeted model improvements. We introduce Matryoshka Transcoders, a novel framework for the automatic discovery and interpretation of physical plausibility features in generative models. Our approach extends the Matryoshka representation learning paradigm to transcoder architectures, enabling hierarchical sparse feature learning at multiple granularity levels. By training on intermediate representations from a physical plausibility classifier and leveraging large multimodal models for interpretation, our method identifies diverse physics-related failure modes without manual feature engineering, achieving superior feature relevance and feature accuracy compared to existing approaches. We utilize the discovered visual patterns to establish a benchmark for evaluating physical plausibility in generative models. Our analysis of eight state-of-the-art generative models provides valuable insights into how these models fail to follow physical constraints, paving the way for further model improvements.

1. Introduction

Recent advances in generative models have achieved remarkable progress in producing photorealistic and instruction-following images [1, 2]. However, these models frequently generate images that appear visually convincing yet violate fundamental physical laws—producing anatomical impossibilities, distorted structures, and malformed text. Such physical plausibility failures significantly undermine the reliability and usability of generative models in high-stakes applications and motivate the development of methods that can identify these failure modes.

Despite their prevalence and importance, physical plausibility errors remain challenging to detect systematically. Existing evaluation metrics primarily focus on semantic alignment between images and text [3–5] or aggregate distribution quality [6, 7], leaving specific physical violations largely undetected [8, 9]. Recent research has shown that even state-of-the-art large multimodal models (LMMs) struggle to consistently identify subtle physical implausibilities when evaluating images holistically [10], motivating us to develop methods that can identify such failure modes of generative models. This detection gap leaves such failure modes largely unidentified, hindering the understanding of how some generated images fail in physical plausibility.

Furthermore, no framework exists for automatically identifying and interpreting specific physical error patterns in natural language. Though recent work [10] has proposed to explore physical plausibility detection by the identification of topic-related visual patterns grounded in natural language, they utilize preliminary interpretability modules and struggle to discover fine-grained error patterns. The absence of interpretable, automatically discovered features means that practitioners lack systematic understanding of the diverse ways generative models violate physical constraints, limiting the development of targeted mitigation strategies.

We introduce Matryoshka Transcoders (See Figure 1), a novel framework for the automatic discovery and interpretation of physical plausibility features in generative models. Our approach utilizes the transcoder architecture [11] to identify topic-relevant features and extends the Matryoshka representation learning paradigm [12] to further enhance its

*We will release our code and dataset upon acceptance.

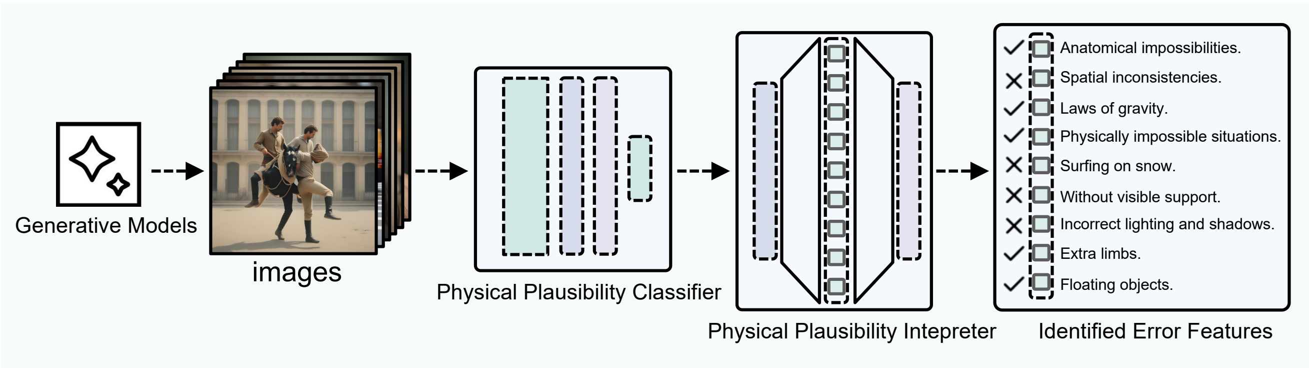

Figure 1. **Matryoshka Transcoders for Automatic Physical Plausibility Analysis in Generative Models.** We pass images from generative models to a physical plausibility classifier, whose representations are analyzed by Matryoshka Transcoders to extract hierarchical sparse features related to physical plausibility. These features encompasses anatomical impossibilities, spatial inconsistencies, gravity violations, and other physical implausibilities and are interpreted by LMMs automatically.

performance. We first assign human annotators to annotate a physical plausibility dataset, including a set of generated images that are marked as physically plausible or not. We use this dataset to train a physical plausibility classifier that contains physical plausibility features, and develop Matryoshka Transcoders on the latent representations of the classifier. The Matryoshka Transcoder identifies specific physical plausibility features, which are then interpreted by LMMs to get their human understandable explanations.

Our experiments demonstrate that Matryoshka Transcoders successfully discover diverse physical plausibility patterns without manual feature engineering, achieving superior feature relevance and accuracy compared to existing methods including standard transcoders, sparse autoencoders, and prior interpretability approaches. We utilize these identified patterns related to physical plausibility to establish the first interpretation-based benchmark for evaluating physical plausibility in generative models. Our analysis of eight state-of-the-art generative models, including SDXL variants [13], DALL-E3 [14], FLUX [15], Kolors [16], and Stable-Cascade [17], reveals systematic failure modes and provides actionable insights into how these models violate physical constraints, from textual distortions to anatomical impossibilities. These findings pave the way for targeted model improvements and more physically plausible generative systems.

Our contributions are listed as below:

- **Targeted Feature Discovery.** We present the first automated pipeline to identify and interpret visual features in ML models relevant to a human-assigned topic.
- **Matryoshka Transcoders.** We extend the Matryoshka paradigm to transcoder architectures, enabling hierarchical sparse feature learning at multiple granularity levels.
- **Feature Relevance Metrics.** We introduce two metrics, *population-based relevance* and *description-based relevance*, to quantitatively evaluate how well discovered features align with human-assigned topics.
- **Benchmarking Physical Plausibility.** We provide the first interpretation-based benchmark for evaluating existing generative models' capability in producing physically plausible images.
- **Insights on Generative Model Failures.** Our approach discovers diverse physical plausibility features, providing valuable insights into how generative models fail.

2. Related Works

2.1. Mechanistic Interpretability

Mechanistic Interpretability [18, 19] is a research field that aims to reverse-engineer neural networks into human-understandable algorithms and concepts. Modern machine learning models represent diverse information in compact representation spaces where individual neurons activate for multiple unrelated concepts—a phenomenon known as polysemanticity—making it difficult to interpret what each neuron represents [20]. A central challenge is decomposing polysemantic neurons within machine learning models into monosemantic neurons that activate for specific interpretable features [21]. To address this, Sparse Dictionary Learning [22, 23] has emerged as a powerful approach for automatically discovering interpretable features in neural network activations. Sparse Autoencoders (SAEs) [24–26] and their variations [27–29] reconstruct neural representations with high sparsity constraints to extract interpretable features. Matryoshka SAEs [12] further improve upon SAEs by simultaneously training nested dictionaries of increasing size, enabling hierarchical feature learning at multiple granularities. Beyond activation reconstruction, Transcoders [11] build sparse mappings between layer activations to identify layer-specific features and have been applied to discover physical plausibility features [10].

2.2. Physical Plausibility in Generative Models

Although recent generative models are remarkably capable of producing instruction-following and realistic outputs [1, 2], they remain prone to notable physical plausibility failures [30, 31]. Existing evaluation approaches primarily focus on semantic alignment or aggregate distribution quality [3–7], leaving specific physical violations largely undetected and unaddressed [8, 9, 32, 33]. Recently, Language-Grounded Sparse Encoders (LanSE) [10] proposed to identify interpretable visual patterns for physical plausibility assessment by imitating human cognition [34–39]. However, this approach suffers from limited feature accuracy and struggles to identify specific erroneous patterns. Our work addresses these limitations by leveraging hierarchical transcoder features trained on physical plausibility classifiers, enabling more accurate and fine-grained identification of diverse physical violations.

3. Method

3.1. Overview

Our method consists of three stages to automatically identify and interpret physical plausibility features in generated images. First, we train a binary classifier on human-annotated images to distinguish between physically plausible and implausible content, using a frozen CLIP vision encoder with a lightweight classification head (§3.2). Second, we extract intermediate activations from this classifier and train Matryoshka Transcoders to learn hierarchical sparse features that capture physical plausibility-relevant patterns at multiple levels of granularity (§3.3). Finally, we automatically interpret the learned features using a two-stage prompting strategy with large multimodal models: first identifying the common visual pattern among top-activating images, then analyzing whether this pattern represents a physical plausibility violation (§3.4). This pipeline enables scalable discovery and interpretation of diverse physical plausibility errors without manual feature engineering.

3.2. Physical Plausibility Classifier Training

We employ two independent human annotators to label a dataset of generated images according to their human understanding on the physical plausibility of these images. Each image is assigned a binary label indicating whether it contains physical plausibility errors or is physically plausible. To ensure annotation quality and consistency, we provide annotators with clear guidelines defining physical plausibility violations and include representative examples of common error types. The annotated dataset is augmented with images from natural image datasets, including MSCOCO [40] and Flickr8k [41], which serve as additional negative samples of physically plausible images.

Utilizing this human-annotated dataset, we train a binary classifier to classify natural images as physically plausible or physically implausible. We use CLIP-ViT-Large-patch14 [42] as the base encoder and attach a two-layer classification head to it. The CLIP vision encoder remains frozen during training to preserve its rich semantic representations learned from large-scale vision-language pretraining. This trained classifier serves as the foundation for extracting intermediate representations that contain physical plausibility information, which will be subsequently analyzed by our Matryoshka Transcoders.

3.3. Matryoshka Transcoders

We extract three layers of activations from the trained classifier: base model activations $\mathbf{h}_2 \in \mathbb{R}^{d_2}$, hidden layer activations $\mathbf{h}_1 \in \mathbb{R}^{d_1}$, and classification outputs $\mathbf{h}_0 \in \mathbb{R}^{d_0}$. We train Matryoshka Transcoders on the transformation from $\mathbf{h}_2$ to $\mathbf{h}_1$ to identify physical plausibility-relevant patterns encoded in the hidden representations.

Model architecture. Our Matryoshka Transcoder is a two-layer model with a sparse latent space, consisting of an encoder layer that maps input activations to sparse features:

$$\mathbf{z} = \text{Enc}(\mathbf{h}_2) = \text{ReLU}(W_{\text{enc}}\mathbf{h}_2 + \mathbf{b}_{\text{enc}}) \tag{1}$$

and a decoder layer that reconstructs the target activations:

$$\bar{\mathbf{h}}_1 = \text{Dec}(\mathbf{z}) = W_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{dec}} \tag{2}$$

where $W_{\text{enc}} \in \mathbb{R}^{d_z \times d_2}$ and $W_{\text{dec}} \in \mathbb{R}^{d_1 \times d_z}$ are the encoder and decoder weight matrices, and d_z is the latent dimension.

Training objective. Our key innovation is to adapt the Matryoshka training paradigm for transcoders. We define a nested sequence of dictionary sizes $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$ where $m_1 < m_2 < \cdots < m_k = d_z$. For each size $m \in \mathcal{M}$, we enforce sparsity by selecting only the top-m activated latents and compute a reconstruction using these features. The total loss aggregates objectives across all granularity levels:

$$\mathcal{L} = \sum_{m \in \mathcal{M}} \|\mathbf{h}_1 - (W_{\text{dec}}\mathbf{z}_{0:m} + \mathbf{b}_{\text{dec}})\|_2^2 \tag{3}$$

where $\mathbf{z}_{0:m}$ denotes the first m latent features, and $W_{0:m}^{\text{dec}}$ are the corresponding decoder weights.

3.4. LMM Interpretation

After training the Matryoshka Transcoders, we extract the top-activating images for each learned feature by passing the entire dataset through the trained model. For each feature, we collect the images that pass one activation threshold, along with their associated text captions. To automatically interpret these features in natural language, we employ a two-stage prompting strategy using LMMs.

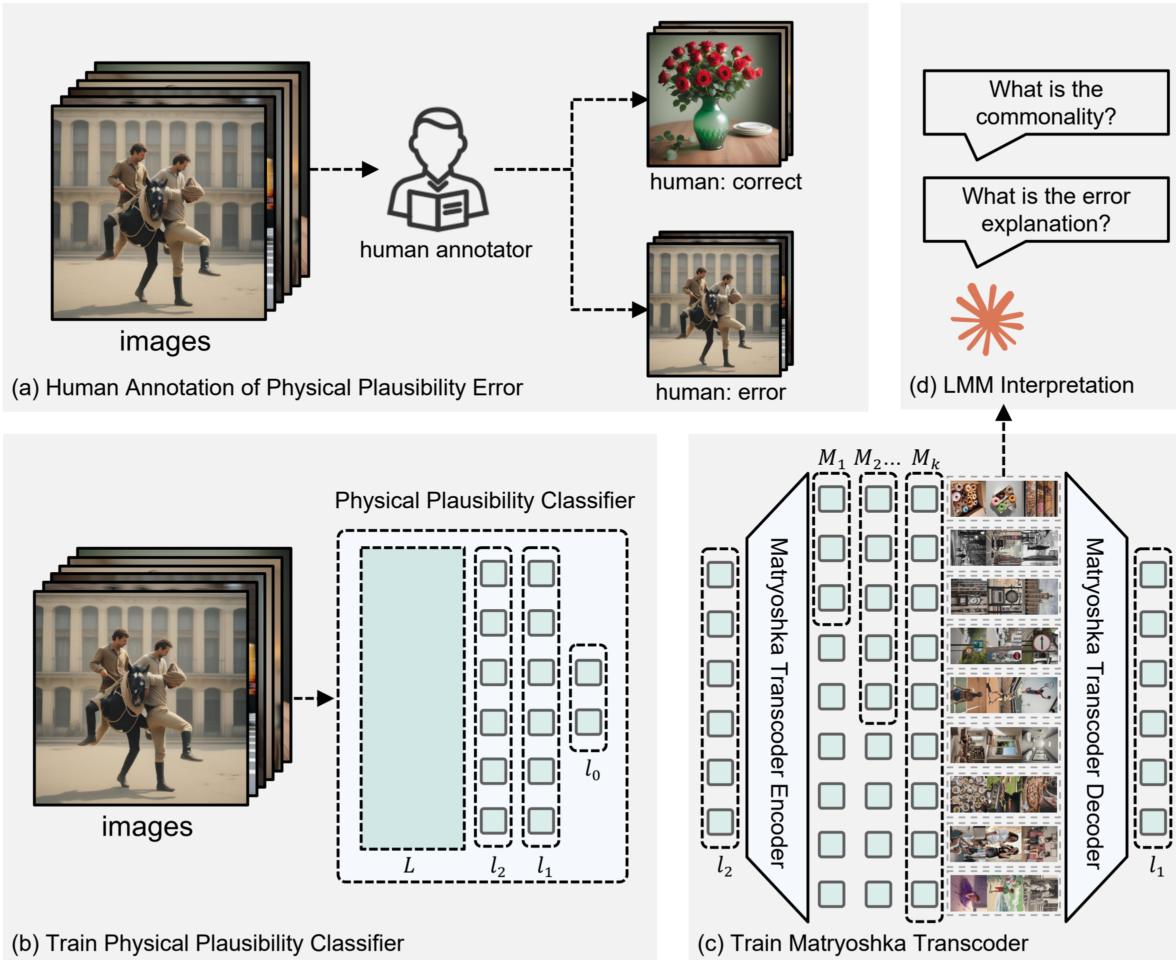

Figure 2. **Complete pipeline for our proposed method, Matryoshka Transcoders.** Our approach combines four stages: (a) human annotation of correct/error images, (b) training a physical plausibility classifier that possess relevant information, (c) training Matryoshka transcoders to discover sparse, 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.

Stage 1: Pattern summarization. We first identify the common visual pattern captured by each feature. For each feature, we provide the LMM with the image-text pairs and prompt it to summarize the shared visual characteristics.

Pattern Summarization Prompt ($\text{Prompt}_{\text{sum}}$)

You are an expert in visual pattern analysis. Given the following images and their captions: {image-text pairs}, analyze the commonalities among them. Concisely summarize one common visual pattern that is present in all instances.

Stage 2: Physical plausibility analysis. Given the identified summarization, we prompt the LMM to determine whether this pattern represents a physical plausibility violation and, if so, to describe the violation in natural language.

Error Interpretation Prompt ($\text{Prompt}_{\text{interp}}$)

You are an expert in physics and physical reasoning. Here are a list of images and their captions with one common theme, {pattern from Stage 1}: {image-text pairs} analyze whether these images possess any visual patterns that violate physical plausibility. Consider the following aspects:
- Anatomic impossibilities (e.g., six fingers in one hand)
- Distorted structures (e.g., highly distorted and unclear facial expressions)

If yes, provide a concise natural language description of such violation possessed by all samples.

This two-stage approach enables us to automatically generate interpretable descriptions for each feature learned by the Matryoshka Transcoder.

4. Experiments

We evaluate Matryoshka Transcoders on their ability to automatically discover interpretable features relevant to physical plausibility and assess their utility for benchmarking generative models. We begin by describing our experimental setup in Section 4.1, then present qualitative and quantitative analysis of discovered features in Section 4.2, demonstrate the application of our method to benchmark eight generative models in Section 4.3, and conclude with ablation studies validating our design choices in Section 4.4.

4.1. Experimental Setup

Dataset construction. We employ human annotators to label generated images as physically plausible or implausible, following clear annotation guidelines for common violation types. Our human-annotated dataset comprises 3,410 images: 888 physically plausible images and 2,522 images containing physical plausibility errors. This dataset is augmented with 5,000 images from MSCOCO [40] as additional negative examples of physically plausible images to train our physical plausibility classifier.

Model configuration. Our classifier uses frozen CLIP-ViT-Large-patch14 [42] with a two-layer head (768→256→1), trained with AdamW at learning rate 10^{-4}. Matryoshka Transcoders map CLIP embeddings (768-dim) to hidden activations (256-dim) with nested dictionaries $\mathcal{M} = \{128, 256, 512, 1024, 2048\}$ using top-k sparsities of $\{16, 32, 64, 128, 256\}$ for each granularity.

Feature interpretation. We use Claude Sonnet 4.5 [43] for automatic feature interpretation via our two-stage prompting strategy (§3.4), extracting top-20 activating images per feature. The same LMM evaluates feature quality by assessing relevance and accuracy.

Computational requirements. Classifier and Matryoshka Transcoder training each require approximately 1 hour on a single NVIDIA A100 GPU. Feature interpretation via API completes within 20 hours.

4.2. Physical Plausibility Pattern Identification

We evaluate the interpretability and relevance of features discovered by Matryoshka Transcoders through both qualitative and quantitative analysis. Our results demonstrate that Matryoshka Transcoders can successfully identifies diverse physical plausibility failure modes and attach highly accurate natural language labels to these failure modes.

4.2.1. Qualitative results.

Figure 3 presents representative discovered features with their automatically generated interpretations. Our method identifies diverse violation patterns in varied error types: Feature 104 captures anatomical distortions in tennis scenes, Feature 243 detects malformed text in posters, and Feature 648 recognizes implausible hand-drawn styles. Each feature demonstrates semantic coherence across its activating images, with natural language descriptions characterizing the specific physical plausibility violation.

4.2.2. Quantitative results.

We evaluate feature quality using two complementary relevance metrics that assess both the selectivity of discovered features and the relevance of their description.

Population-based relevance. For each feature i with activation threshold θ, we compute its topic-relevance score (wrong ratio) as:

$$M_i = \frac{|\{x : f_i(x) > \theta \wedge y(x) = \text{error}\}|}{|\{x : f_i(x) > \theta\}|} \tag{4}$$

where $f_i(x)$ is the feature activation on image x and $y(x)$ indicates the error label. This quantifies the fraction of a feature's activating images that contain physical plausibility errors. A feature is classified as topic-relevant if $M_i \geq 0.95$ (see Section 5 for justification), meaning at least 95% of its activating images exhibit physical violations. The population-based relevance score then measures the concentration of such selective features:

$$R_{\text{population}} = \frac{|\{i : M_i \geq 0.95\}|}{|\text{Total Features}|} \tag{5}$$

Description-based relevance. We also evaluate whether the automatically generated natural language descriptions capture topic-relevant patterns. For each feature, we first generate a description using our two-stage LMM interpretation pipeline (Section 3): $\text{description}_i = \text{LMM}(\text{images}_i, \text{Prompt}_{\text{sum}})$. We then assess relevance via:

$$R_{\text{description}} = \frac{|\{i : \text{LMM}(\text{description}_i) = \text{relevant}\}|}{|\text{Total Features}|} \tag{6}$$

where $\text{LMM}(\text{description}_i)$ indicates whether an LMM judges the description as relevant to physical plausibility violations. Together, $R_{\text{population}}$ assesses feature selectivity while $R_{\text{description}}$ evaluates interpretation quality.

We benchmark our method against sparse autoencoders [44], transcoders [11], and Matryoshka SAEs [12] as baselines. Specifically, we utilize a large corpus of images generated by different models with diverse captions to identify samples that activate each feature in these interpretability modules, enabling us to compute their relevance scores. Table 1 shows that Matryoshka Transcoders outperform all baselines on both relevance metrics, achieving 16.00% population-based and 17.52% description-based scores. This demonstrates the effectiveness of combining targeted transcoder learning with hierarchical Matryoshka training for discovering physical plausibility features.

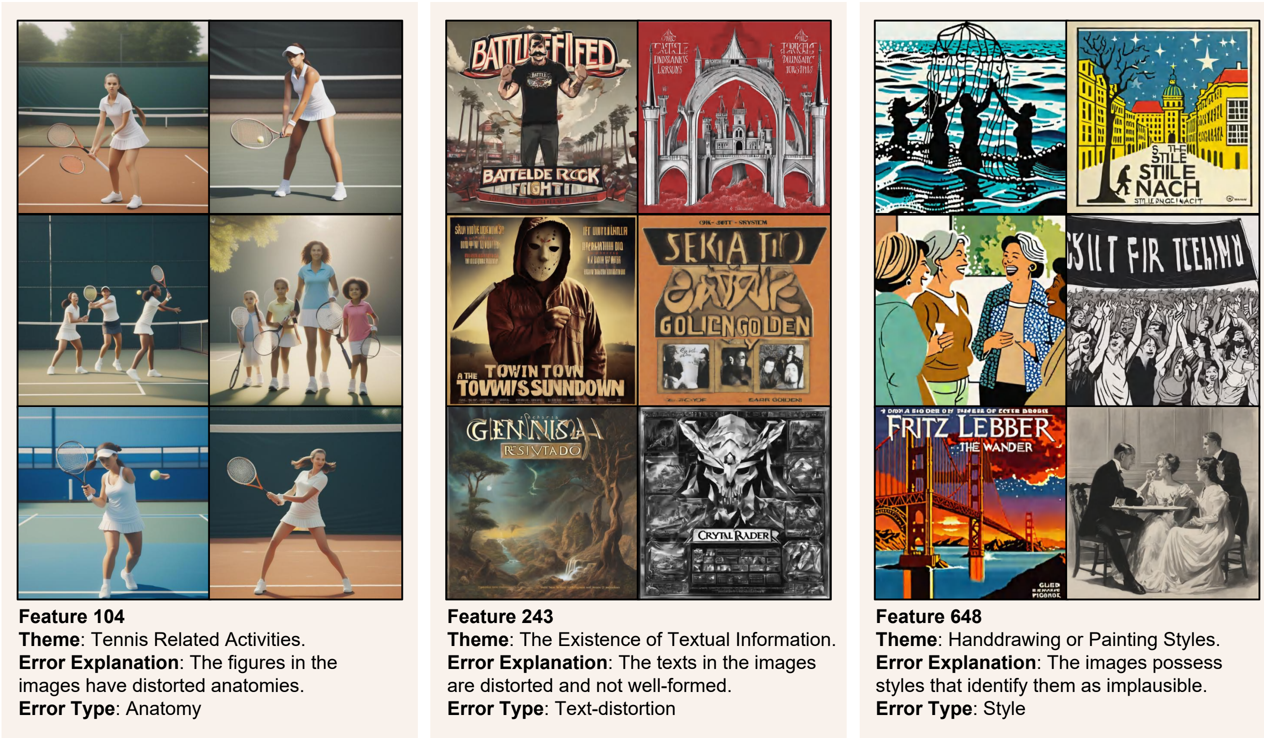

Figure 3. **Qualitative examples of physical plausibility features discovered by Matryoshka Transcoders.** Each column shows a representative feature with its top-activating images, automatically generated theme, error explanation, and categorized error type.

Method	$R_{\text{population}}$	$R_{\text{description}}$
SAE	6.54%	6.01%
Matryoshka SAE	9.88%	5.66%
Transcoder	12.74%	11.52%
Matryoshka Transcoder	**16.00%**	**13.04%**

Table 1. **Main Evaluation Results.** We evaluate methods' capabilities to identify topic-relevant features by comparing their feature relevance scores. Matryoshka Transcoders achieve superior performance on both metrics compared with all other baselines.

4.3. Generative Model Benchmark

We apply discovered physical plausibility features to benchmark eight state-of-the-art generative models. By analyzing which features activate on generated images, we obtain fine-grained diagnostics of each model's physical plausibility failures—revealing not just error frequency but also specific explanations. Table 2 summarizes our findings.

For each generative model, we compute the mean error feature count as:

$$\text{Error Count} = \frac{1}{N}\sum_{j=1}^{N} |\{i : f_i(x_j) > \theta \wedge M_i \geq 0.95\}| \quad (7)$$

where N is the number of generated images, and M_i is the topic-relevance score. This metric counts how many topic-relevant features activate on each image, averaged across all images. Lower scores indicate better performance, as fewer error-selective features respond to the model's outputs.

Model Name	Mean Error Count
SDXL-turbo	74.2
SDXL-base	53.1
DALL-E3	49.8
SDXL-medium	48.7
SDXL-large	46.8
Stable-Cascade	43.1
Kolors	40.8
FLUX	**38.0**

Table 2. **Physical plausibility evaluation of state-of-the-art generative models.** Error count represents the average number of physical plausibility features activated per generated image across 5,000 test samples. Lower values indicate better performances. FLUX achieves the best physical plausibility, while SDXL-turbo exhibits substantially more violations than other models. Image samples of these generative models can be found in Appendix.

4.4. Ablation Studies

We study how the latent dimension size affects our results.

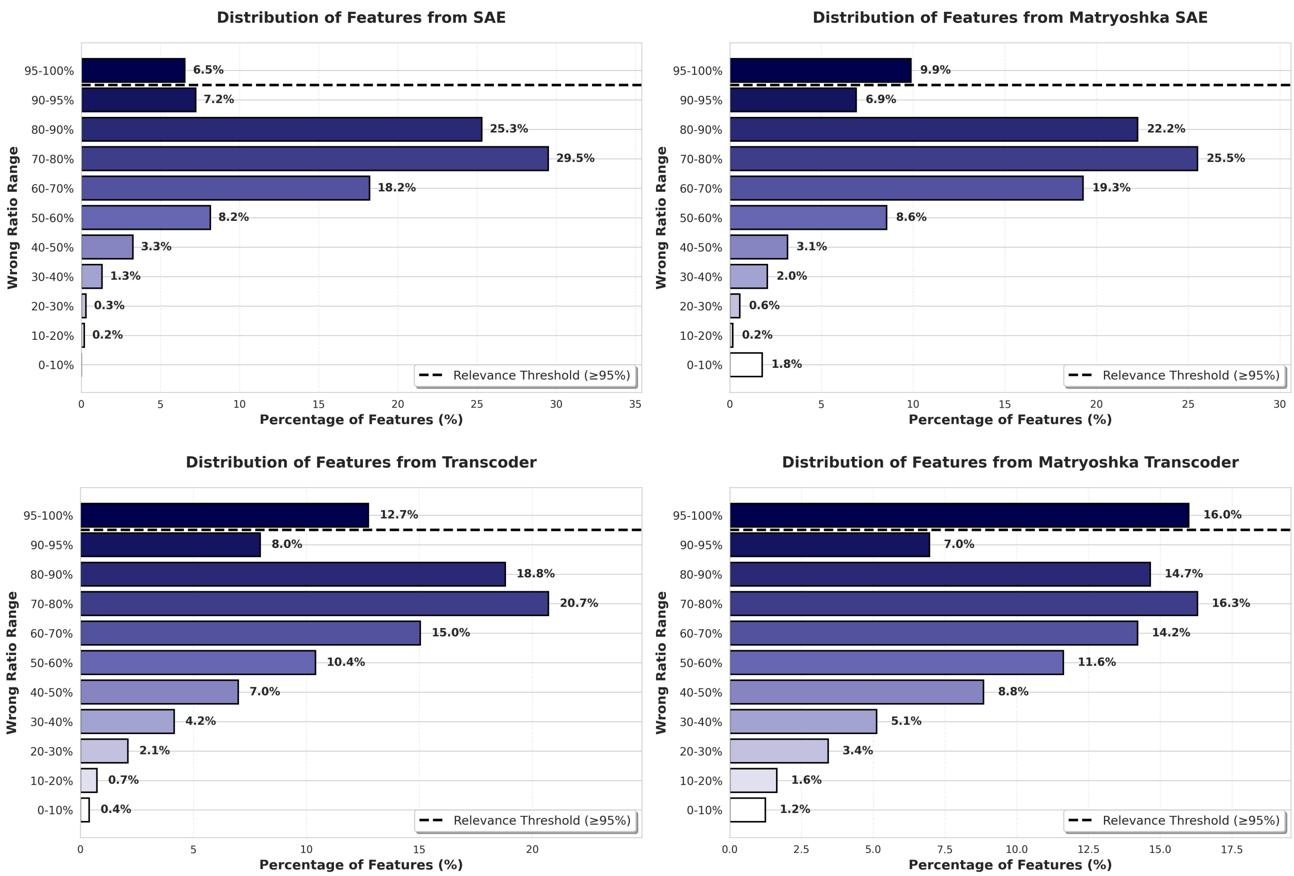

Figure 4. Feature relevance distribution across methods. Percentage distribution of features by their opic-relevance score (wrong ratio) across four sparse dictionary learning methods. Features above the relevance threshold of 0.95 are considered as task-relevant features.

Dictionary size. Figure 5 shows a clear tradeoff: smaller dictionaries (1,024) yield more focused and relevant features (17.2% relevance), while larger ones (16,384) cover more types of errors but with lower relevance (11.6%).

Feature stability. The selectivity of error-related features remains stable (about 67–70%) across all sizes, and most features are active (94–98%), showing efficient use of the dictionary.

In general. Small dictionaries give concentrated, reliable features; large ones give broader coverage. We use nested sizes (M = 128, 256, 512, 1024, 2048) to combine both benefits.

5. Discussion

We introduce Matryoshka transcoders to enable interpretable feature discovery on targeted topics. By training on physical plausibility classifier representations, our method enables automatic discovery of interpretable features relevant to human-assigned topics.

The Feature Relevance Bottleneck Phenomenon. As demonstrated in Figure 4, feature distributions by the topic-relevance metric M (defined in Equation 4) reveal a clear bottleneck: features cluster at two extremes ($70\% < M < 80\%$ and $M > 95\%$) rather than exhibiting a uniform spread or single-peak distribution. This bottleneck indicates a natural boundary between topic-relevant features that selectively respond to topic-specific visual patterns and task-irrelevant features that merely correlate with the presence of physical plausibility errors. We therefore classify features with $M \geq 0.95$ as topic-relevant. Matryoshka architectures enhance this specialization, yielding 9.9% and 16.0% relevant features compared to 6.5% and 12.7% for standard architectures. The bottleneck phenomenon justifies focusing interpretability efforts on highly selective features, as intermediate-relevance features likely represent polysemantic patterns that resist coherent interpretation.

6. Limitations

- **Training integration.** Though our features successfully characterize physical violations, they are not yet incorpo-

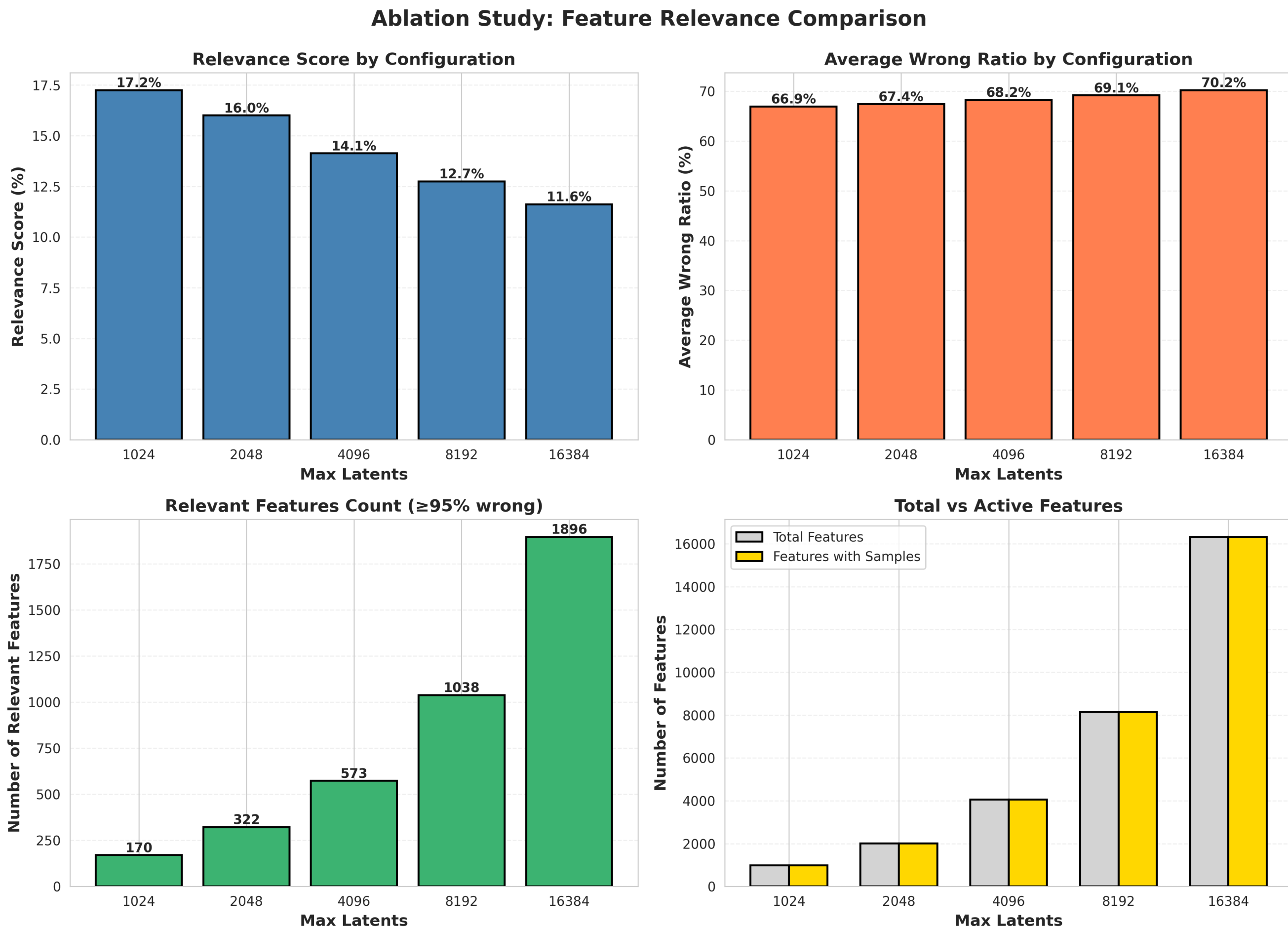

Figure 5. **Effect of Latent Dimension Size on Feature Discovery.** Ablation study evaluating five dictionary sizes (1024-16384 latents). Smaller dictionaries achieve higher relevance scores (17.2% vs 11.6%) but discover fewer total relevant features (170 vs 1,896), revealing a quality-quantity tradeoff. Average wrong ratio remains stable (67-70%), indicating consistent feature selectivity across configurations. Nearly all features remain active (94-98%), showing efficient dictionary utilization regardless of size.

rated into training pipelines to actively reduce such failures, which lay as one promising future work.

- **Feature coverage.** Certain subtle or rare physical violations may be underrepresented if training data lacks sufficient examples of such patterns, potentially creating evaluation blind spots for edge cases.
- **Feature absorption rate.** We do not quantify the feature absorption phenomenon [45] for our methods. Computing absorption rates for visual inputs is substantially more challenging than for language models due to the high-dimensional, continuous nature of image representations. Future work should develop methods to measure feature completeness in vision models.
- **Evaluation scope.** Our framework focuses on physical plausibility, leaving other important quality dimensions, such as aesthetics, semantic faithfulness, and social biases, for future work.
- **Single LMM dependency.** Our feature interpretation pipeline relies on a single large multimodal model, and the generated explanations may vary depending on the choice of LMM. While we find the interpretations to be consistently coherent across features, future work should evaluate interpretation consistency across LMMs.

7. Conclusion

We introduced Matryoshka Transcoders for automatically discovering and interpreting physical plausibility features in generative models. Our hierarchical approach learns sparse features at multiple granularity levels, enabling interpretable evaluation of physical plausibility and systematically revealing failure modes across generative models. Our work demonstrates that targeted feature discovery is achievable through domain-specific classifiers and developing mechanistic understanding on the classifiers.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 3

[2] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024. 1, 3

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 3

[4] Georgia Gabriela Sampaio, Ruixiang Zhang, Shuangfei Zhai, Jiatao Gu, Josh Susskind, Navdeep Jaitly, and Yizhe Zhang. Typescore: A text fidelity metric for text-to-image generative models, 2024.

[5] Ziwei Yao, Ruiping Wang, and Xilin Chen. Hifi-score: Fine-grained image description evaluation with hierarchical parsing graphs. In *European Conference on Computer Vision*, pages 441–458. Springer, 2024. 1

[6] Yu Yu, Weibin Zhang, and Yun Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 3(11), 2021. 1

[7] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024. 1, 3

[8] Lorenzo Bianchi, Fabio Carrara, Nicola Messina, and Fabrizio Falchi. Is clip the main roadblock for fine-grained open-world perception?, 2024. 1, 3

[9] Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models, 2025. 1, 3

[10] Yiming Tang, Arash Lagzian, Srinivas Anumasa, Qiran Zou, Yingtao Zhu, Ye Zhang, Trang Nguyen, Yih-Chung Tham, Ehsan Adeli, Ching-Yu Cheng, Yilun Du, and Dianbo Liu. Human-like content analysis for generative ai with language-grounded sparse encoders, 2025. 1, 2, 3

[11] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits, 2024. 1, 2, 5

[12] Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders, 2025. 1, 2, 5

[13] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. 2

[14] Evgeniy Hristoforu. Dall·e 3. https://huggingface.co/ehristoforu/dalle-3-xl-v2, 2023. Accessed: 2025-03-14. 2

[15] Black Forest Labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. Accessed: 2025-03-14. 2

[16] Kwai-Kolors. Kolors-diffusers. https://huggingface.co/Kwai-Kolors/Kolors-diffusers, 2024. Accessed: 2025-03-14. 2

[17] Stability AI. Stable cascade: Efficient text-to-image generation in highly compressed latent spaces. https://huggingface.co/stabilityai/stable-cascade, 2025. Technical Report. 2

[18] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. 2

[19] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2025. 2

[20] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. 2

[21] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. 2

[22] Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning, 2024. 2

[23] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. 2

[24] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. 2

[25] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[26] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. 2

[27] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024. 2

[28] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024.

[29] Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders, 2024. 2

[30] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. Presented at CVPR 2024, June 2024. Available at `https://doi.org/10.1109/CVPR52733.2024.00239`. 3

[31] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting, 2024. 3

[32] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 3

[33] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 3

[34] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 3

[35] Semir Zeki. Art and the brain. *Journal of Consciousness Studies*, 6(6-7):76–96, 1999.

[36] Sophie J Nightingale, Kimberley A Wade, and Derrick G Watson. Can people identify original and manipulated photos of real-world scenes? *Cognitive research: principles and implications*, 2:1–21, 2017.

[37] Hany Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), 2022.

[38] Paulo E. Santos, Roberto Casati, and Patrick Cavanagh and. Perception, cognition and reasoning about shadows. *Spatial Cognition & Computation*, 18(2):78–85, 2018.

[39] Michael S Landy. Texture analysis and perception. *The new visual neurosciences*, 476:639–652, 2013. 3

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3, 5

[41] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. 3

[42] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training, 2025. 3, 5

[43] Anthropic. Claude 4.5 sonnet. `https://www.anthropic.com/claude`, 2024. date: 2025-11-05. 5

[44] Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014. 5

[45] David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders, 2025. 8

A. Qualitative Examples of the Generative Model We Evaluate

To complement our quantitative evaluation in Table 2, we present qualitative samples from all eight evaluated generative models in Figure 6. For each model, we selected a diverse set of representative images spanning common generation scenarios. These examples validate that our metric captures meaningful differences in physical plausibility that align with human perception of image quality.

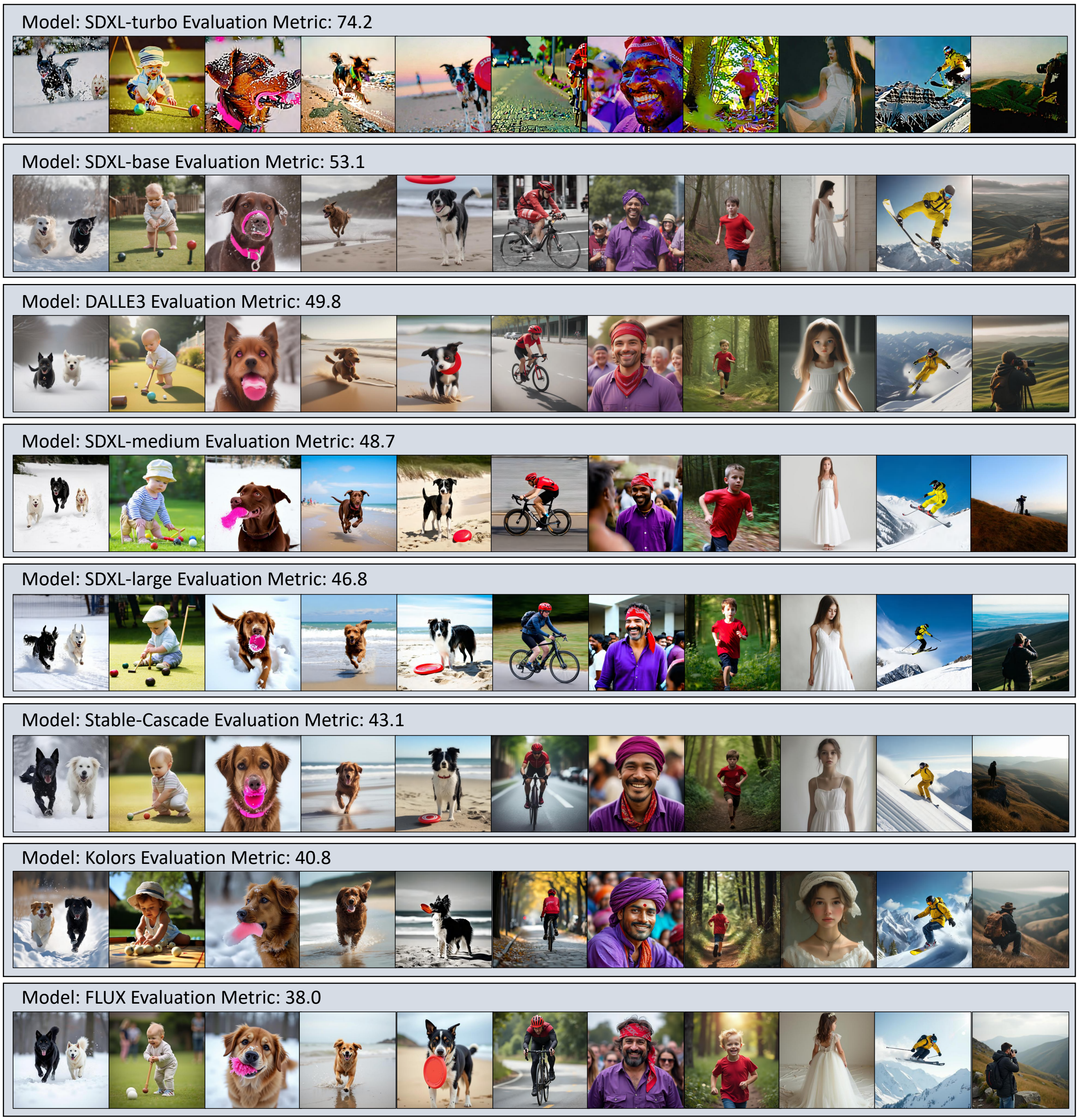

Figure 6. **Qualitative samples from evaluated generative models ranked by physical plausibility.** Each row shows representative outputs from a different model, ordered by mean error count. Lower error counts correspond to better visual quality and fewer physical violations, with FLUX achieving the best performance (38.0) and SDXL-turbo exhibiting the most plausibility failures (74.2).

B. Full LMM Prompts

In this section, we fully describe the prompts we use in our pipelines in detail.

Pattern Summarization Prompt ($\text{Prompt}_{\text{sum}}$)

You are an expert in multimodal feature analysis.
Given the following images and their captions: {image-text pairs}.
Analyze the commonalities among these images. Identify: If there exist one common feature that is possessed by all the instances.
Summarize and output exactly one feature, for example: '[Commonality: Animal wildlife in natural habitats]' and '[Commonality: Strawberry-based dessert or dish]'.
Only answer in general, do not analyze each image one by one, only generate one single concise phrase. Start answer with '[Commonality:'. End with ']'

Error Interpretation Prompt ($\text{Prompt}_{\text{interp}}$)

You are an expert in analyzing visual content for physical plausibility errors, anatomical accuracy, and generation artifacts in AI-generated images.
I will show you$len(samples)$ sample images from a learned feature in an interpretability module trained to detect physical plausibility errors.
Feature Commonality: "$commonality$" **Error Ratio**: $error_{ratio}$ of these images contain physical errors ($error_{count}/total_{count}$). Be strict with the evaluations. Usually only error ratios larger than 0.95 indicate clear physical plausibility failure mode as these features should not be activated in the presence of any correct images. Other features can be unusable and should be considered as no such error mode.
Here are the sample images:
{image-text pairs}.
Based on the $len(samples)$ images shown above and knowing that: - Feature commonality: "commonality" - $error_{ratio}$ contain physical plausibility errors
Analyze whether these images share a **common physical plausibility error or anatomical inaccuracy**. Focus ONLY on:
Physical Errors: - Incorrect number of fingers, toes, or limbs - Extra or missing body parts - Distorted anatomy or impossible body proportions - Unnatural poses or joint configurations - Incorrect object physics (floating, defying gravity) - Impossible spatial arrangements or perspectives - Anatomically incorrect faces or features - Object inconsistencies or impossible constructions
Important: - Ignore style, artistic choices, or image quality - Ignore semantic content unless it relates to physical errors - Focus ONLY on violations of physical or anatomical plausibility
Response Format: If there IS a common physical error across these images: [Error: Brief description of the specific error]
If there are NO clear physical plausibility errors for all images or if the error ratio does not suggest clear and monosemantic error pattern: [No common errors]
Your response MUST start with either "[Error:" or "[No common errors]" and end with "]".

C. Effect of Dictionary Size on Feature Distribution.

Figure 7 demonstrates how dictionary size affects the quality and quantity of discovered physical plausibility features. We trained Matryoshka Transcoders with five nested dictionary sizes ranging from 1,024 to 16,384 latents and analyzed the distribution of features by their topic-relevance scores. Smaller dictionaries (1,024 latents) achieve higher feature relevance scores (17.2%), producing more concentrated and reliable features, but discover fewer total relevant patterns (170 features). In contrast, larger dictionaries (16,384 latents) provide broader coverage of error types, discovering 1,896 relevant features, but with lower average relevance (11.6%). Notably, all configurations exhibit a consistent bimodal distribution with features clustering at 70-80% and above 95% relevance, revealing a natural bottleneck that separates task-correlated features from truly topic-specific features. The average wrong ratio remains stable (67-70%) across all dictionary sizes, and nearly all features remain active (94-98%), indicating efficient dictionary utilization. This analysis motivates our nested training approach with $\mathcal{M} = \{128, 256, 512, 1024, 2048\}$, which combines the high relevance of small dictionaries with the comprehensive coverage of larger ones.

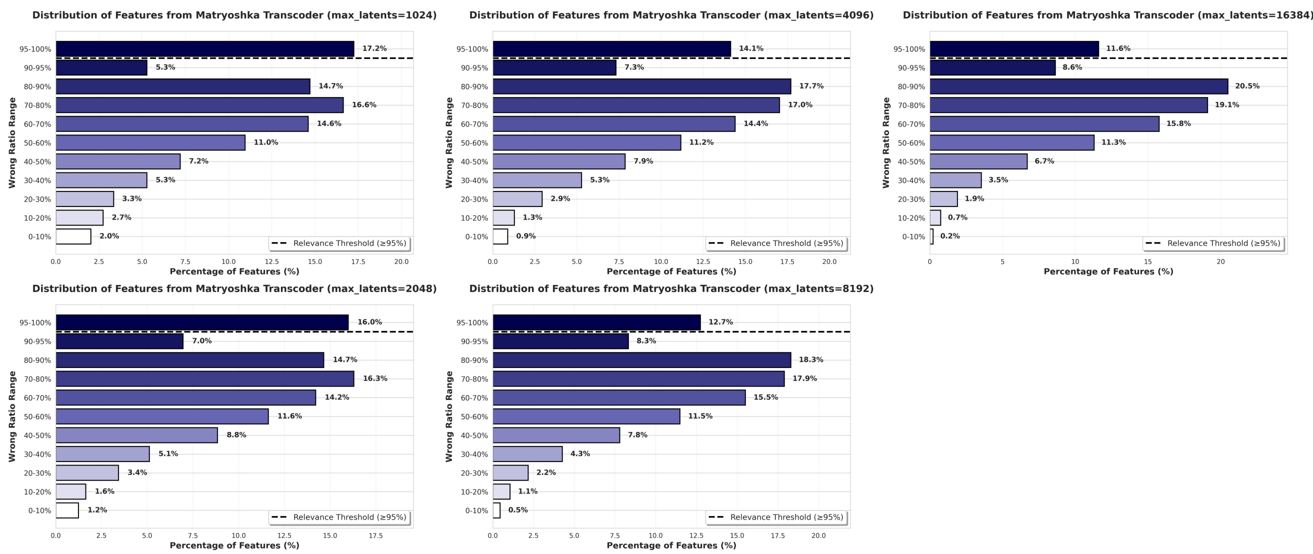

Figure 7. **Feature relevance distribution across dictionary sizes in Matryoshka Transcoders.** Distribution of features by their topic-relevance score (wrong ratio) for five nested dictionary sizes (1024, 4096, 8192, 16384 latents). Features above the relevance threshold of 0.95 (indicated by dashed line) are considered topic-relevant. Smaller dictionaries yield higher concentration of relevant features (17.2% for 1024) but discover fewer total patterns, while larger dictionaries achieve broader coverage (1,896 relevant features for 16384) with lower relevance scores (11.6%). The bimodal distribution consistently shows feature clustering at $70-80\%$ and $\geq 95\%$ relevance ranges across all configurations, revealing a natural bottleneck between task-correlated and topic-specific features.