GRADE: A Fine-grained Approach to Measure Sample Diversity in Text-to-Image Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Existing diversity metrics like Fréchet Inception Distance (FID) and Recall require reference images and are generally not reliable. Evaluating the diversity of text-to-image (T2I) model outputs remains a challenge, especially in capturing fine-grained variations essential for creativity and bias mitigation. We propose **Gr**anular **A**ttribute **Di**versity **E**valuation (GRADE), a descriptive and fine-grained method for assessing sample diversity in T2I models without requiring reference images. GRADE estimates the distribution of attributes within generated images of a concept, such as the shape or flavor distribution of the concept "cookie", and computes its normalized entropy, providing interpretable insights into model behavior and a diversity score. We show GRADE achieves over 90% agreement with human evaluation while having weak correlation to FID and Recall, indicating it captures new, fine-grained forms of diversity.

We use GRADE to measure and compare the diversity of 12 T2I models and reveal that the most advanced models are the least diverse, scoring just 0.47 entropy and defaulting to depicting concepts with the same attributes (e.g., cookies are round) 88% of the time, despite varied prompts. We observe an inherent trade-off between diversity and prompt adherence, akin to the Precision-Recall trade-off and negative correlation between diversity and model size. We identify that underspecified captions in training data contribute significantly to low sample diversity, leading models to depicting concepts with the same attributes. GRADE serves as a valuable tool for benchmarking and guiding the development of more diverse T2I models.

1 Introduction

Text-to-image (T2I) models have the remarkable ability to generate realistic images based on textual descriptions. However, textual prompts are inherently *underspecified* [1, 2], meaning they cannot fully describe all visual attributes that appear in the resulting image. Often, we aim for our T2I models to *produce diverse outputs* that represent the full spectrum of possible visual attributes. For example, when sampling generations of "a cookie", we expect to see cookies with different ingredients, colors, and backgrounds, among other variations. But are current T2I models truly capable of generating diverse outputs? Evaluating diversity is inherently challenging because the set of possible visual attributes—the support set—is virtually infinite. Standard metrics, such as Fréchet Inception Distance (FID) [3] and Recall [4, 5], are limited in their ability to capture a human-like understanding of diversity. In this paper, we introduce a novel method for evaluating diversity. Our approach leverages language models (LMs) and visual question-answering (VQA) models to approximate the support set of visual attributes for a given concept. Once this set is established, we use the VQA model to automatically quantify the frequency of each visual attribute, and we assess diversity using entropy or other diversity measures.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Figure 1: A random sample of 4 images from each prompt for the concept "cookie" by FLUX.1-dev. The prompt does not describe the cookie in a particular way, yet, all cookies are round with the vast majority having chocolate chip on top.

We propose **Gr**anular **A**ttribute **D**iversity **E**valuation (GRADE), a fully automated method for measuring sample diversity in T2I models at a fine-grained level, focusing on attributes of concepts, such as the shape attribute of the concept cookie. Existing diversity evaluation methods predominantly rely on comparing the generated output to a set of *reference images*, which are assumed to represent the diversity inherent in the concept of interest. However, this approach has two key limitations. First, reference images are often unavailable. More critically, the set of possible visual attributes for any concept is virtually infinite. While diversity can be measured in terms of pixel color distribution, a human understanding of diversity is rooted in the *semantics* of the image: the relevant attributes are highly specific to each concept, and identifying them requires extensive world knowledge. For example, when generating an image of a parachute, we might be interested in attributes such as its color, orientation, or whether it is open or closed. Other potential details, like the exact number of clouds in the background, may be irrelevant. Existing diversity metrics fail to capture the set of meaningful visual attributes for each concept. In contrast, we leverage the world knowledge embedded in visual question-answering (VQA) models to approximate this set, offering a scalable approach that generalizes well to new concepts.

Our approach (Fig. 4) involves using an LM to generate prompts to elicit diverse outputs from T2I models and questions pertaining to visual attributes of each concept. We then use a VQA to extract information about the visual attributes using the images and questions. We translate the answers to a categorical distribution by using an LM to generate plausible categorical answers to each question and map the answers to categories. Each category expressed in natural language, and its diversity quantifiable using normalized entropy. For instance, the shape distribution of cookie by FLUX.1-dev [6] is {round: 0.983, flower: 0.015, star: 0.0016}) and its normalized entropy is 0.03 bits.

Using GRADE, we show T2I models exhibit *default behaviors*, consistently generating certain attributes for given concepts. For instance, the prompt "a cookie in a bakery" results in round cookies for over 98% of the time for the models we tested, neglecting other possible shapes like square or rectangular, as shown in Fig. 1. This bias can stem from imbalanced training data or the models' tendency to optimize for the most probable outputs.

Our contributions are threefold:

- **A Novel Diversity Evaluation Method:** We introduce GRADE, a fine-grained and descriptive method for evaluating sample diversity in T2I models without the need for reference images. We show GRADE measures diversity better than FID and Recall, even with the presence of reference images.
- **Comprehensive Diversity Analysis:** Using GRADE, we conduct an extensive study comparing the diversity of 12 T2I models, revealing that even the most diverse ones do not achieve high diversity. Our analysis uncovers *hints* of a trade-off between prompt adherence and diversity, akin to the Precision-Recall trade-off [5] and negative correlation between model size and diversity.
- **Insights into Training Data Influence:** We demonstrate that underspecified captions in training data contribute to low diversity.

2 Related Work

We provide background on previous metrics like Fréchet Inception Distance (FID) [3] and Precision-and-Recall [4, 5] in Appendix A.1.

Most evaluation use-cases fall into one of three: (1) to gauge how well the model learned the training set; (2) to compare model performance; and (3) to evaluate a model on a particular concept.

Existing approaches, like FID, Precision-and-Recall, and others compare a distribution of generated images to a distribution of reference images–a set of images that exhibits the desired level of diversity. This can be the model's training data, to measure how well it covers it, or an established dataset, like ImageNet [7], which models can then use to compare.

However, testing performance on a new class is non-trivial. Especially if the class is fine-grained, such as an attribute of a concept. For example, to evaluate the model's ability to generate the shape attribute for the cookie concept, we would need to collect images that depict cookies in various shapes and ensure the style of images is consistent with the style in the model's training set, as this can affect the diversity score. This approach is not designed to measure the diversity of specific concepts within the images evaluation, but overall similarity between two distributions. In general, distribution-comparison metrics do not directly evaluate the *diversity* of the model, but its *fit* to some data.

We take a different approach: we define diversity over the attributes of concepts, use GRADE to estimate categorical distributions of concepts' attributes, and compare them to uniform behavior using entropy. Our approach covers the three popular use-cases: (1) GRADE naturally works with new concepts and attributes, without collecting data, as we detail in Section 3; (2) GRADE can compare the diversity between models, as we demonstrate in Section 4; and (3) GRADE can estimate distributions from the dataset and compare them to distributions in inference time, as we show in Appendix G.

3 GRADE: Measuring T2I Diversity

Approach. We measure diversity in T2I models by measuring the relation between concepts (e.g., "cookie" or "helmet") and their attributes (e.g., "shape" or "color"). Each concept c, is an object that can be described textually and visualized graphically, and an attribute $\mathcal{A}_c$ is a set of characteristics or states of c that can vary among different images. We analyze the relation between them over prompts that are underspecified, i.e., that mention c but not $\mathcal{A}_c$. We model their relation as a categorical distribution, such that every value $a \in \mathcal{A}_c$ (e.g., round $\in A_{shape}$) is a possible category in $\mathcal{A}_c$. The distribution can then be estimated on prompts that do not specify the attribute (e.g., "a cookie in a bakery" does not specify a shape). We show a sample of concepts, attributes, and categories in Table 1.

After estimating the distribution, we can quantify its entropy. We use *normalized entropy*, henceforth referred to as entropy: $H_N(X) = \frac{H(X)}{\log_2(|\mathcal{A}_c|)}$, where H is the entropy of the distribution, $|\mathcal{A}_c|$ is the support size of the distribution, and its range is from 0 to 1. When close to 1, the attribute is almost uniform, and when close to 0, it is almost always the same category.

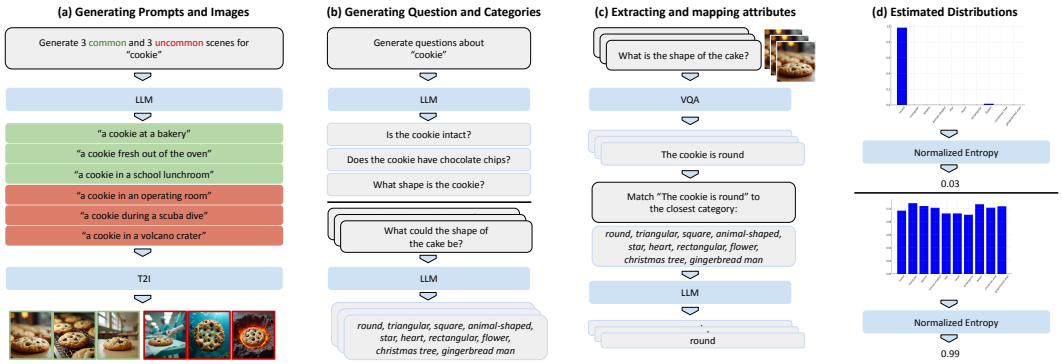


Figure 2: An overview of GRADE with "cookie" as the input concept c and the output is distributions $P(c, \mathcal{A}_c)$. The figure shows four steps: (a) the prompt and image generation. (b) Question and category generation. Questions dictate the attributes we test for diversity ($\mathcal{A}_c$). Each question defines a distribution $P(c, \mathcal{A}_c)$. For convenience, the figure shows the workflow with a single distribution. (c) Attribute extraction. Questions are first answered with a VQA. Then, an LLM maps the answers to matching categories. (d) Estimate the diversity with entropy.

Each distribution $P(c, \mathcal{A}_c)$ is estimated using several prompts that mention c but not its attribute $\mathcal{A}_c$. We say that this is a *concept distribution*, as it models the relation between c and $\mathcal{A}_c$ across more than one prompt. While these are our primary subject in this work, we also report results on each *prompt distribution* $P(p_c, \mathcal{A}_c)$, which models the relation between c and $\mathcal{A}_c$ over the prompt p that mentions c but not $\mathcal{A}_c$.

We arrive at a representative measure of the overall diversity of the model by measuring the entropy over various distributions across different concepts and attributes, and compute the mean entropy.

3.1 Method

We measure the diversity of T2I models over concepts, we design GRADE, a pipeline that estimates the diversity of a concept c from end-to-end: It generates various prompts for each concept c, to invoke varied responses from a T2I model, which are then processed to estimate $P(c, \mathcal{A}_c)$. Fig. 4 illustrates GRADE, Appendix D details each step.

4 Comparing Diversity of T2I Models

Having GRADE at our disposal, we turn to measure and compare the diversity ability of T2I models to generate diverse instances of concepts. To achieve our goal, we use GRADE to estimate attribute distributions of concepts to measure the diversity of T2I models. We first provide an overview of our concepts and distributions, then describe the models, and finally the results.

Table 1: Concepts, attributes, and categories. Each (concept, attribute) pair makes for a distribution. For readability, we show only one attribute per concept with more examples in Appendix C.

Concept	Question (Attribute)	Categories
Bin	What material is the bin made from?	mesh, cardboard, carbon fiber, rubber, wood, bamboo, wicker, plastic, ceramic, stainless steel, fiberglass, metal, aluminum, steel, fabric, glass
Person	Does the person appear to be alone or with others?	alone, with others
Suitcase	Is this a vintage suitcase?	yes, no
Cake	What flavor is the cake?	tiramisu, cheesecake, carrot, chocolate, strawberry, vanilla
Pool	Is the pool indoor or outdoor?	indoor pool, outdoor pool

4.1 Experimental Setup

Data and distributions overview. For each model, we estimate distributions over 100 common concepts such as "cake" and "suitcase". Each concept is linked to 4 questions on average. In total, there are 400 concept distributions and 2400 prompt distributions, consisting a total of 60,000 images per model.

T2I models. We measure the diversity of the 12 models shown in Table 2. All models were used with the default hyperparameters.

Table 2: The 12 T2I Models grouped by family ygmove citations to model column.

Family	Model
Stable Diffusion	SD-1.1, SD-1.4, and SD-2.1 [8], SDXL [9], SDXL-Turbo [10], SDXL-LCM [11], SD-3 (2B) [12]
IF-DeepFloyd	DeepFloyd-M, DeepFloyd-L, and DeepFloyd-XL [13]
FLUX	FLUX.1-schnell [14], FLUX.1-dev [15]

4.2 Results

Table 3 shows the mean entropy of models across all concept and prompt distributions. No model exhibits high diversity scores. As expected, the diversity of prompt distributions is even lower compared to concept distributions, since the images were only generated from a single prompt.

Surprisingly, FLUX.1-dev is the least diverse model, despite its impressive capabilities and the statements that the model was trained with output diversity in mind [6]. One explanation for this disparity is that some concepts were targeted for data variation while others received less attention. Appendix I illustrates the diversity.

Diversity in relation to model scale. Entropy over both concept and prompt distributions demonstrate decline almost in tandem with increased model size, especially within each family of models we test, as shown in Appendix E. This *hints* toward an *inverse-scale law* [16] between model size and entropy. However, this is not to claim that more parameters results in lower diversity, but that they are correlated. An alternative hypothesis is that larger models are fine-tuned for prompt adherence on top of their pretraining. Fig. 3 demonstrates a tradeoff between prompt adherence and sample diversity, reminiscent to the tradeoff between Recall and Precision [5]. Plots based on the prompt distributions are included in Appendix E.

Default behaviors. The outputs of GRADE are naturally descriptive. Both the outputs by the VQA and the categories are in natural language and can be used to explain the resulting entropy. Here, we focus on the latter and observe that the distributions we approximate are often heavily skewed toward a certain category. When a single category is 80% likely, it is a *default behavior* of the model, but any value indicating a single category the majority of the time is appropriate. We define this phenomenon over both types of distributions. In Appendix F we show a sample of the phenomenon and that almost all models exhibit at least one default behavior over every concept, with SD-1.1 exhibiting 87.13% prompt distributions, with a similar trend over the concept distributions.

5 Conclusion

We introduced GRADE, a fully automated fine-grained evaluation method for measuring sample diversity in T2I models, based on concepts and their attributes. We achieve over 90% agreement with human evaluations while showing weak correlation with traditional metrics like FID and Recall, indicating that we capture fine-grained forms of diversity not reflected by existing measures.

Using GRADE, we conduct a comprehensive analysis of 12 state-of-the-art T2I models and uncovered a prevalent limitation: these models default to generating images with the same attributes for a concept on anywhere from 78% to 90% of the concepts we tested, with an increasing trend as models scale and improve in prompt adherence. This default behavior highlights a significant gap in current models'

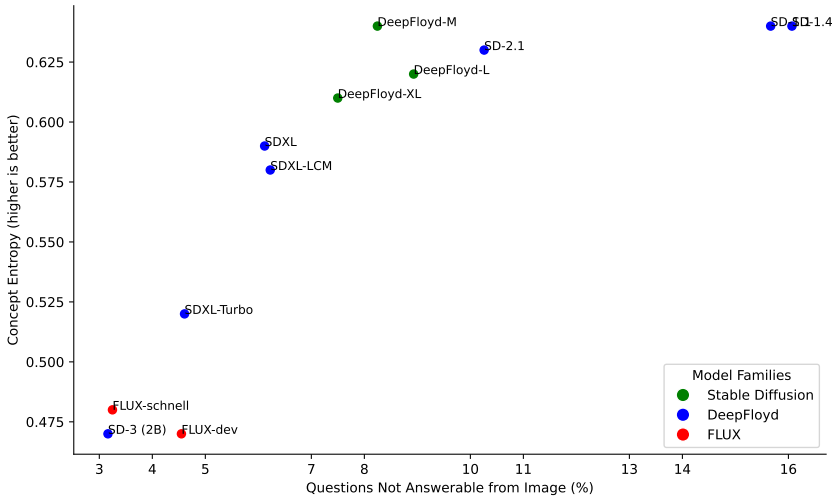


Figure 3: The mean concept entropy of the models plotted against the % of "none of the above". The plot shows there is strong negative correlation between diversity and prompt-adherence, which hints at a tradeoff. The x-axis accounts for cases where the concept c is not depicted in the prompt that mentions it (low prompt adherence) or there is no category in $\mathcal{A}_c$ that matches the answer. In practice, around 80% of the cases indicate the latter.

Table 3: **Entropy in concept- and prompt-level distributions.** The mean entropy over all distributions for each model in both setups. Values close to 1 indicate highly diverse behavior (uniform) while values close to 0 indicate highly repetitive categories. The *most* diverse models are in bold.

	Mean Entropy ↑	
Model	**Concept-level**	**Prompt-level**
DeepFloyd-M	**0.64**	0.49
DeepFloyd-L	0.62	0.47
DeepFloyd-XL	0.61	0.46
SD-1.1	**0.64**	**0.54**
SD-1.4	**0.64**	0.53
SD-2.1	0.63	0.51
SDXL	0.59	0.46
SDXL-Turbo	0.52	0.36
SDXL-LCM	0.58	0.45
SD-3 (2B)	0.47	0.34
FLUX.1-schnell	0.48	0.36
FLUX.1-dev	0.47	0.32

ability to capture the rich diversity inherent in visual concepts and suggests that specialized prompt adherence fine-tuning is at odds with sample diversity.

Our investigation into training data reveals that underspecified captions contribute significantly to low sample diversity, leading models to adopt default behaviors. This finding underscores the importance of detailed and representative training data in enhancing diversity.

By offering a descriptive metric for diversity, GRADE serves as a valuable tool for benchmarking T2I models and guiding their development toward greater diversity. Future work could explore methods to enrich training data, incorporate diversity-promoting mechanisms during model training, and extend GRADE to evaluate relationships between different attributes of a concept (e.g., the relationship between the shape and suggested flavor distributions of "cookie"). Ultimately, we hope that our work will inspire more nuanced evaluations and drive advancements in generating diverse visual content from textual descriptions.

References

[1] Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. Underspecification in scene description-to-depiction tasks. *arXiv preprint arXiv:2210.05815*, 2022.

[2] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: flaws in word-to-concept mapping in text2image models. *arXiv preprint arXiv:2210.10606*, 2022.

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[4] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[5] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

[6] Black Forest Labs. Announcing black forest labs. `https://blackforestlabs.ai/announcing-black-forest-labs/`, August 2024. Accessed: 2024-08-29.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[9] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[10] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[11] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[13] DeepFloyd Lab at StabilityAI. DeepFloyd IF: a novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding. `https://www.deepfloyd.ai/deepfloyd-if`, 2023. Retrieved on 2023-11-08.

[14] Black Forest Labs. FLUX.1-dev Model Documentation. `https://huggingface.co/black-forest-labs/FLUX.1-schnell`, 2024. Accessed: Aug 24 2024.

[15] Black Forest Labs. FLUX.1-dev Model Documentation. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024. Accessed: Aug 24 2024.

[16] Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023.

[17] Dongkyun Kim, Mingi Kwon, and Youngjung Uh. Attribute based interpretable evaluation metrics for generative models. *arXiv preprint arXiv:2310.17261*, 2023.

[18] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke

Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[19] OpenAI. Introducing structured outputs in the api, 2024. Accessed: 2024-09-17.

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models, 2022.

[21] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data?, 2024.

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, D. Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2014.

[23] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA, 2013. Association for Computing Machinery.

A Appendix

Compute. All the experiments detailed in this paper can be run on a single A100-80GB. We used up to 4 at a time.

GPT-4. We use `gpt-4-turbo-2024-04-09` with max tokens set to 512 and temperature set to 0.

GPT-4o. We use `gpt-4o-2024-08-06` with the structured ouputs setting, max tokens set to $1,000$, and temperature set to 0.

A.1 Related Work: Extended

Background. Fréchet Inception Distance (FID) [3] measures fidelity and diversity by calculating the distance between them. The result is a score that represents both fidelity and diversity, but does not reflect the trade-off between them. Precision-and-Recall is designed to separate them [4] and [5] proposed a variation that makes the trade-off between them explicit. Recently, [17] proposed two methods to measure similarity between generated images to images from a training set over predefined attributes. Both assume a training set of images and a predefined set of text-based attributes reflected in them. The first measures the difference in density for each attribute and the second extends this measurement to joint-attribute relationships (e.g., is the correlation between "male" and "sunglasses" similarly strong as in the training set?).

B LM Prompting

B.1 Concept Collection

To collect a list of diverse concepts, we prompt GPT-4 [18] with the following:

```
Provide a CSV of 100 unique concepts, like the example below.
concept_id is an enumeration that begins from 0.
Choose concepts that are easy to visually verify for a VQA model.

concept_id,concept
0, an ice cream
1, a cake
2, a suitcase
3, a clock
```

B.2 Prompt generation

We generate our prompts using the following template. The following prompt was used to generate typical prompts:

```
Please suggest three typical settings for the concept below.
Note that the output should be a list of strings.

Here's an example:
Concept: a cake
Prompts: [
"a cake in a bakery,
"a cake at a birthday party",
"a cake at a swimming pool"
]

Concept: {concept}
```

While this one was used to generate atypical prompts:

```
Please suggest three atypical settings for the concept below.
Note that the output should be a list of strings.

Here's an example:
Concept: a cake
Prompts: [
"a cake on a weight loss clinic,
```

```
"a cake at a gym",
"a cake at a swimming pool"
]

Concept: {concept}
```

B.3 Question generation

To extract attribute values, we collect questions automatically and in an unsupervised fashion. We prompt GPT-4 to analyze the attributes of the concept provided in the prompt and use it as context to generate questions.

```
Help me ask questions about images that depict certain concepts.

I will provide you a concept.
Your job is to analyze the concept's typical attributes
and ask simple questions that can be answered by viewing the image.

Here's an example:

concept:
a cake
attributes:
cakes can be made in different flavors, shapes,
and can have multiple tiers.

questions:
1. Is the cake eaten?
2. Does the cake have multiple tiers?
3. In what flavor is the cake?
4. What is the shape of the cake?
5. Does the cake show any signs of fruit on the outside or
suggest a fruit flavor?

Now that you understand, let's begin.

concept: {c}
```

B.4 Category generation

To define $\mathcal{A}_c$ and generate categories, we provide GPT-4 [18] with a concept, a question, and a prompt. GPT-4 then outputs a list of categories that can match the question. The process is performed for all prompts associated with the concept. The sets are then unified with similar answers removed (e.g., "motorbike helmets" is removed, because "motorcycle helmets" already exists). The result of the unification is $\mathcal{A}_c$.

```
I have a question that is asked about an image.
I will provide you with the question and a caption of the image.
Your job is to analyze the description of the image and the question,
hypothesize plausible answers that can surface from viewing the image.
Then, I need you to list the plausible answers.

For example,
    Caption: a helmet in a bike
    Question: What type of helmet is depicted in the image?

    Plausible answers:
    answers = ["motorcycle helmets",
               "bicycle helmets",
```

```
                "football helmets",
                "construction helmets",
                "military helmets",
                "firefighter helmets",
                "rock climbing helmets",
                "hockey helmets"]
Now your turn.
    Caption: {c}
    Question: {q}
    Plausible answers:
```

B.5 Generating answers

We use GPT-4o to answer the generated questions with $1,000$ as max tokens and temperature 0. Our prompt is straightforward:

```
Answer the following question with one of the categories. To come up with the
↪  correct answer, carefully analyze the image and think step-by-step before
↪  providing the final answer.

Question: {question}
Categories:{categories}
Selection:
```

C Extended Data Overview

Table 4: Concepts and their attributes with detailed distributions.

Concept	**Question (Attribute)**	**Attribute Values**
Bin	What shape is the bin?	circular, octagonal, square, cylindrical, triangular, rectangular, round, oval, hexagonal
Bin	Does the bin have a lid?	yes, no
Person	Is the person male or female?	male, female
Person	Does the image show the person from up-close?	yes, no
Suitcase	Is the suitcase open or closed?	open, closed
Suitcase	Is the suitcase soft-shell or hard-shell?	soft-shell suitcase, hard-shell suitcase
Cake	Does the cake have multiple tiers?	yes, no
Cake	Is the cake eaten?	yes, no
Pool	Is there anyone swimming in the pool?	yes, no
Pool	What color is the water in the pool?	reflective like a mirror, black, clear, green, blue, brown

D GRADE: Measuring T2I Diversity Extended

D.1 Method

We measure the diversity of T2I models over concepts, we design GRADE, a pipeline that estimates the diversity of a concept c from end-to-end: It generates various prompts for each concept c, to invoke varied responses from a T2I model, which are then processed to estimate $P(c, \mathcal{A}_c)$. Fig. 4 illustrates GRADE, below we detail each step.

Table 5: A random sample of generated prompts, their type (typical or atypical), Nouns, and the number of captions that include all nouns in LAION-2B.

Prompt	Type	Nouns	Frequency
a drawer in an office	typical	"drawer", "office"	51,560
hiking boots a mountain trail	typical	"hiking", "boots", "mountain", "trail"	327
a clown in a circus	typical	"clown", "circus"	39,950
a crown in a costume shop	typical	"crown", "costume", "shop"	66
a mailbox outside a house	typical	"mailbox", "house"	6,188
a banana in a blimp	atypical	"banana", "blimp"	2
a hat in a jewelry store	atypical	"hat", "jewelry", "store"	156
a bin on a mountain peak	atypical	"bin", "mountain", "peak"	5
a cow in a desert oasis	atypical	"cow", "desert", "oasis"	0
popcorn in a refrigerator	atypical	"popcorn", "refrigerator"	64

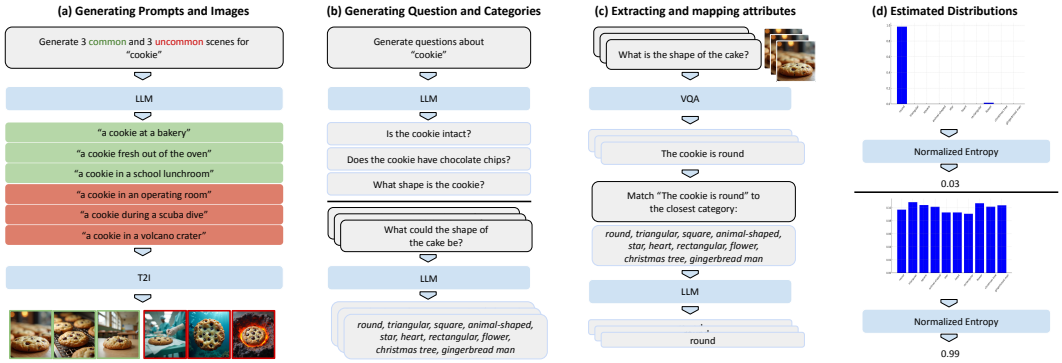


Figure 4: An overview of GRADE with "cookie" as the input concept c and the output is distributions $P(c, \mathcal{A}_c)$. The figure shows four steps: (a) the prompt and image generation. (b) Question and category generation. Questions dictate the attributes we test for diversity ($\mathcal{A}_c$). Each question defines a distribution $P(c, \mathcal{A}_c)$. For convenience, the figure shows the workflow with a single distribution. (c) Attribute extraction. Questions are first answered with a VQA. Then, an LLM maps the answers to matching categories. (d) Estimate the diversity with entropy.

(a) Generating images of c. To measure the model's ability to generate attributes of c, GRADE generates two types of prompts for each concept: (1) prompts describing the concept in common contexts, like "a cookie fresh out of the oven" (common prompts); (2) prompts describing the concept in uncommon contexts, such as "a cookie in an operating room" (uncommon prompts). The former is likely congruent with the training data, and may surface prevalent associations between the concept and certain categories that mesh the prompt. The latter places the concept in a scene that it is less likely to have been observed in the training set. This creates a disentangling effect from associations to categories the prompt might introduce. If a category is common across both common and uncommon prompts it is ingrained to the concept, independent from a specific prompt. For example, cookies are often generated round.

(b) Generating attributes and categories. After generating prompts and images, GRADE generates questions that probe for information about attributes of concepts. Each question pertains to an attribute $\mathcal{A}_c$. We provide an LM with a concept and instruct it to first outline its attributes, then generate questions that can be answered simply by viewing an image of the concept. For example, upon receiving the input "cookie", our LM noted that cookies can be made in different shapes, and proposed the question "What is the shape of the cookie?". Next, we need categories to map the natural language answers returning from the VQA. We use an LM to generate plausible categories to the questions given the prompts. The categories are then unified to a set. For example, the question "What is the shape of the cookie?" is paired with the set A_{cookie} = {rectangular, round, square, ...}.

(c) Extracting attributes. To analyze images for attributes, GRADE pairs all images and questions associated with the same concept and feeds it to a VQA, which outputs an answer in natural language. GRADE then provides an LLM with an answer and the set of categories pertaining to the distribution and instructs it to select the closest matching category (e.g., "The cookie is round" will be mapped to the "round" category). Answers that are mapped to "none of the above" are discarded. We end up with frequency distributions and normalize each to cumulatively sum to 1.

(d) Estimated distributions. GRADE outputs discrete probability distributions. Every (concept, attribute) pair results in a concept distribution. Every (prompt, attribute) pair results in a prompt distribution. We use the probability structure to compute the entropy and output a diversity score from 0 to 1.

Implementation details. In step (a) we generate 3 prompts from each type and 100 images using each prompt. In step (b) we generate an average of 4 questions (which represent attributes) per concept. We note that these values can be modified. The exact hyperparameters and prompts are detailed in Appendix A and Appendix B.

Throughout the steps above, we use GPT-4 [18] as our LLM, and while we present the question answering and category mapping as separate steps, we do them in one-shot using GPT-4o and the Structured Outputs feature [19]. We validate GRADE with human evaluation in Appendix D.2 and compare GRADE to previous metrics in Appendix D.3

D.2 Validating GRADE

GRADE is a modular system; as such, we validate the quality of each module separately.

Prompt generation. We review the generated prompts to ensure they reflect their categories, common or uncommon. We extract the nouns from each prompt and check their co-occurrence in LAION-5B [20] using WIMBD [21], a tool that allows to count and search large datasets efficiently. We find that on average, the co-occurrence of the common categories is 30,655, while for the uncommon, it is 956.

Question generation. We validate that indeed all questions can be answered just from viewing an image that faithfully depicts the generated prompts.

VQA alignment with human evaluation. We validate the quality of GPT-4o using Amazon Mechanical Turk (AMT). Each example includes a question, an image, and the plausible categories (including the "none of the above" option). The workers are requested to select the category that best matches the question and image. We take the majority decision over three submissions. Visuals of the task with further detail on the task are provided in Appendix H.

We run this evaluation twice. First to test the overall quality of GPT-4o, using a sample of 1,000 images generated by our models and a second time to demonstrate we can recreate the distributions estimated with GRADE by replacing GPT-4o with workers. Specifically, we run the task with all 600 images of the "what is the shape of the cookie?" concept distribution for three models: SD-1.4, SDXL-Turbo, and FLUX.1-dev.

In the first scenario, GPT-4o aligns with the majority decision for **90.2%** of the time and in the second, GPT-4o aligns with the majority decision **92.8%** of the time: SD-1.4 88% of the time, FLUX.1-dev 91.2%, and SDXL-Turbo 99.5%.

Plausible category generation. We verify the coverage of the categories we generate by examining "none of the above" selections by *either* majority decisions or the VQA for the 1,000 examples we uploaded to AMT. In total, there are 115 selections. Out of these, only 3 are because the answer is not reflected in one of the categories. In 92 times, the T2I model did not include the concept mentioned in the prompt, and in the other 20, the VQA or workers did not answer the question correctly. We provide a sample of cases in Appendix H.

D.3 GRADE Compared to Previous Metrics

FID [3] and Recall [5] are two of the most popular diversity measurements, however, they suffer from many issues, such as ... []. We demonstrate that GRADE is weakly correlated to them, which taken

Table 6: **PCC between distribution-based metrics.** Each SD model is compared against its dataset. FID and Recall show low to moderate degrees of correlation across models, while the TVD based on the distributions from GRADE exhibits weak correlations with both. This indicates the distributions estimated by GRADE capture diversity existing metrics do not.

Model	Dataset	FID vs Recall	TVD vs FID	TVD vs Recall
SD-1.1	LAION-2B	-0.41	+0.14	+0.04
SD-1.4	LAION-2B	-0.48	+0.18	-0.10
SD-2.1	LAION-5B	-0.12	-0.16	-0.15

together with the 90% VQA-human agreement reported in Appendix D.2, indicates GRADE captures forms of diversity these metrics do not.

To measure the correlation, we compare images from the LAION datasets to images generated by Stable Diffusion (SD) models, which were trained on LAION. Specifically, we use the setup in Appendix G, where we estimate the same 50 random concept distributions twice. Once by sampling relevant images from LAION and once by generating images using the captions linked to the images we sampled from LAION. For example, if an image we sampled from LAION in a is linked to the caption "Unicorn Cookie", the second distribution will contain an image that was generated using that caption as a prompt. Each caption is used to generate an image. We end up with 50 pairs of training-set and generated images distributions that should closely match. Each distribution consisting of 115 images.

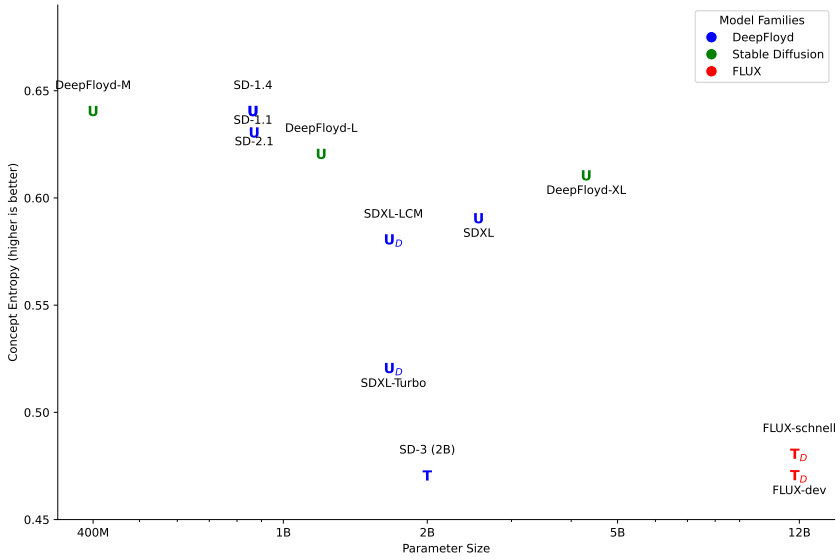
We compute the FID and Recall, using Inception v3 [22] for feature extraction (64 dimensions), and $k = 3$ for Recall.

We compare GRADE with these metrics, using Total Variation Distance (TVD) instead of entropy, as it naturally matches the requirement of FID and Recall for reference images and thus allows for a fair comparison.

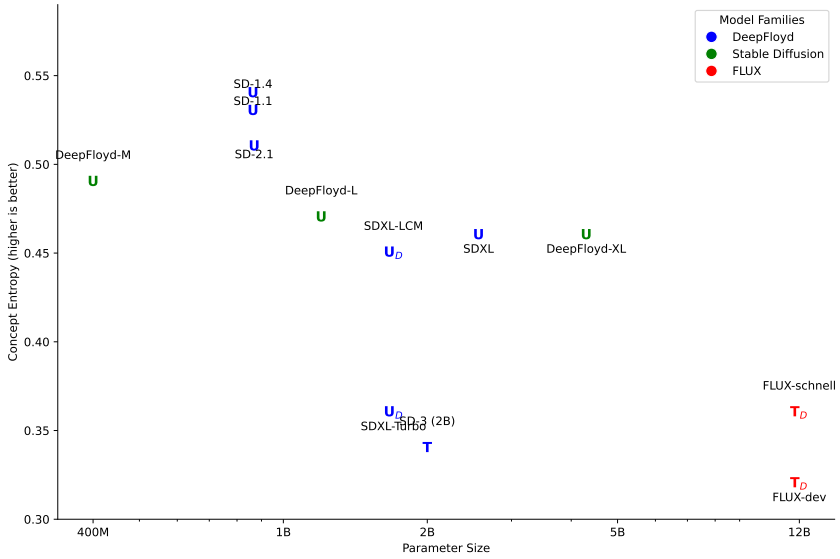
The Pearson Correlation Coefficient (PCC) between FID, Recall, and TVD is shown in Table 6. Both metrics are weakly correlated to GRADE, which achieved over 90% agreement with human evaluation in Appendix D.2. The results imply that FID and Recall do not accurately capture fine-grained similarities such as the attributes of a concept.

E Extended Results

The figures showing plots of diversity against model scale can be viewed in Fig. 5 and the plots of diversity against non-answerable questions in Fig. 6.

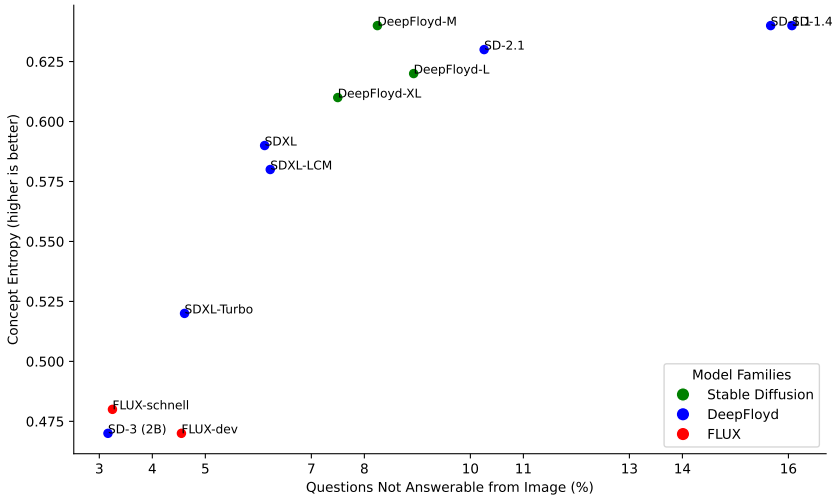


(a) The mean concept entropy of the models potted against the denoiser's parameter size.

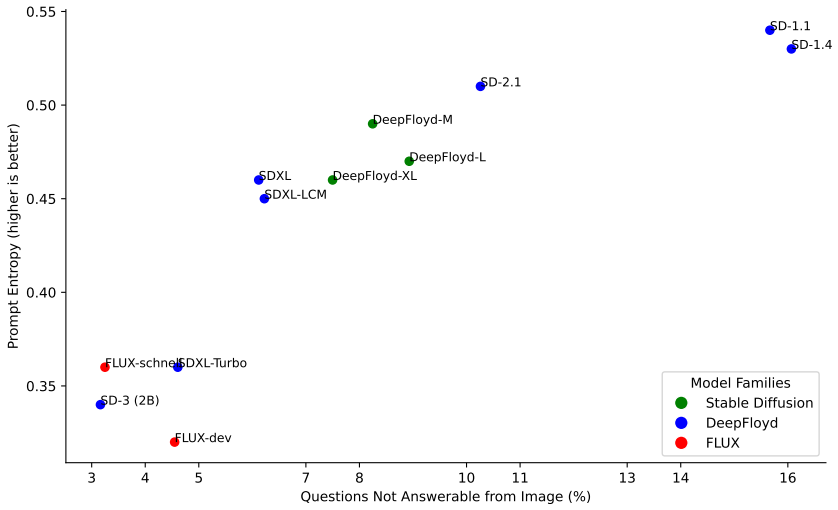


(b) The mean prompt entropy of the models plotted against the denoiser's parameter size.

Figure 5: (a) The mean concept entropy of the models plotted against the denoiser's parameter size. (b) The mean prompt entropy of the models plotted against the denoiser's parameter size. Models marked with U denote U-Net-based models, T denotes transformer-based models. U_D and T_D denote distilled models. To a degree, diversity deteriorates in tandem with parameter size. This phenomenon effect is most apparent within every model family.



(a) The mean concept entropy of the models plotted against the % of "none of the above".



(b) The mean prompt entropy of the models plotted against the % of "none of the above".

Figure 6: (a) The mean concept entropy of the models plotted against the % of "none of the above". (b) The mean prompt entropy of the models plotted against the % of "none of the above". This plot shows there is strong negative correlation between diversity and prompt adherence, which hints at a tradeoff. The x-axis accounts for cases where the concept c is not depicted in the prompt that mentions it (low prompt adherence) or there is no category in $\mathcal{A}_c$ that matches the answer. In practice, around 80% of the cases (see Appendix D.2) indicate the latter.

Table 7: **Percentage of at least one default behavior.** Almost all concepts are associated with at least one default behavior in prompt distributions, with a similar trend in concept distributions.

Model	% of Default Behavior ↓	
	Concept-level	**Prompt-level**
SD-1.1	78	87.13
SD-1.4	82	87.11
SD-2.1	76	89.11
SDXL	81	90.76
SD-3 (2B)	88	94.88
FLUX.1-schnell	**90**	**96.53**
FLUX.1-dev	88	95.71
SDXL-Turbo	86	95.21
SDXL-LCM	82	92.24
DeepFloyd-M	83	91.72
DeepFloyd-L	81	91.75
DeepFloyd-XL	80	91.42

F Default Behaviors: Extended Results

In Section 4.2 we define default behaviors and mention that almost all concepts are associated with at least one default behavior, with SD-1.1 exhibiting 87.13% prompt distributions, with a similar trend over the concept distributions, as shown in Table 7. In Table 8, we report the total number of default behaviors for both types of distributions. Table 9 shows a sample of default behaviors detected in concept distributions and Fig. 7 images of these behaviors.

Table 8: **Percentage of all default behaviors.** There are 100 concept and 2400 prompt distributions in total. The table quantifies the total percentage of default behaviors observed. The *most* diverse models are in bold.

Model	% of Default Behavior ↓	
	Concept-level	**Prompt-level**
SD-1.1	**38.67**	**49.36**
SD-1.4	40.15	50.51
SD-2.1	39.66	51.89
SDXL	44.09	56.97
SD-3 (2B)	55.67	69.41
FLUX.1-schnell	55.17	67.10
FLUX.1-dev	55.67	70.17
SDXL-Turbo	49.75	67.34
SDXL-LCM	43.60	57.10
DeepFloyd-M	38.92	54.07
DeepFloyd-L	39.41	56.10
DeepFloyd-XL	39.90	55.68

Table 9: **A random sample of default behaviors.** The concept is underlined in the question column. Images corresponding to the behaviors in the table can be viewed in Fig. 7.

Model	Question (Attribute)	Category	Percentage
SD-1.1	Is the brick alone or in a stack with others?	stacked	97.4
SD-1.4	Is there a frame around the mirror?	yes	92.9
SD-2.1	Is the suitcase soft-shell or hard-shell?	hard-shell	88.3
SDXL	Is the detective female or male?	male	99.6
SD-3 (2B)	Is the tie a necktie or a bowtie?	necktie	100
FLUX.1-schnell	Is the clock analog or digital?	analog	100

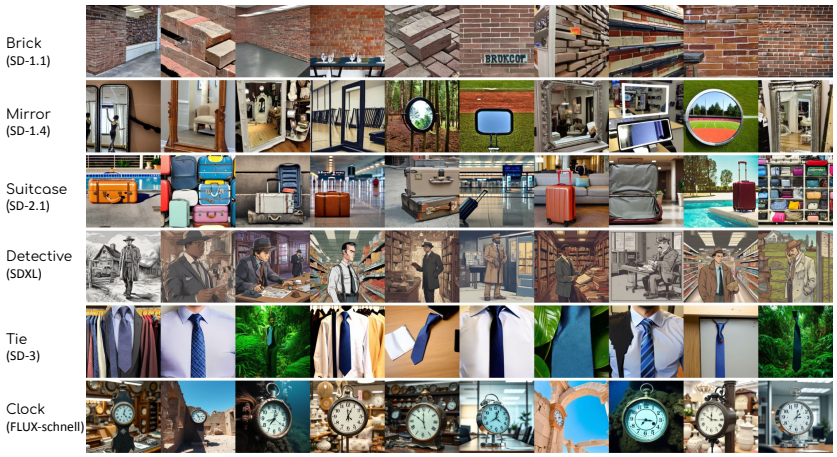


Figure 7: **A sample of images depicting the default behaviors in Table 9.** The concept is shown in the left column with the model directly below it. Images were sampled randomly from all prompts. The default behaviors, top down: (1) stacked bricks; (2) framed mirrors; (3) hard-shell suitcase; (3) male detective; (4) neckties; and (5) analog clocks.

Table 10: **Similarities between images by SD and its training set.** The LAION datasets exhibit moderate diversity based on entropy values. Generating images using prompts from LAION results in comparable entropy and high PCC (strong correlation). Additionally, the small TVD values suggest that the distributions are structurally similar. This pattern remains largely consistent when using generated prompts, with only a slight reduction in correlation and distribution similarity.

Model	Dataset	Source of Prompts	Model Entropy	Dataset Entropy	PCC	TVD
SD-1.1	LAION-2B	LAION-2B	0.62	0.64	0.86	0.11
		Generated	0.58		0.71	0.18
SD-1.4	LAION-2B	LAION-2B	0.62	0.64	0.88	0.10
		Generated	0.60		0.72	0.17
SD-2.1	LAION-5B	LAION-5B	0.68	0.65	0.73	0.13
		Generated	0.68		0.61	0.18

G Is Low Diversity Rooted in the Training Data?

In the previous section we show that all T2I models we experiment with exhibit low diversity scores when attributes are underspecified in the prompt. We hypothesize this behavior can be explained from the pretraining data used to train these models. In this section we explore this hypothesis.

We hypothesize that lack of diversity originates from *reporting bias* [23]: the phenomenon where the obvious is rarely stated. In our case, we hypothesize that captions mentioning concepts seldom specify their common attributes, although they are depicted in the image. For example, it learned to predict "round cookie" by training on images that mention only "cookie", but show it round. Using LAION, the only open-source model that also released its training data, we estimate 50 under-specified concept distributions and compare them to those generated by models trained on LAION. We compare SD-1.1 and SD-1.4 to LAION-2B and SD-2.1 to LAION-5B [20].

Estimating Distributions from LAION We are interested in estimating the diversity of training images whose captions satisfy two conditions: (1) mention the concept and not its periphery (e.g., it should mention a cookie and not cookie cutter) and (2) do not imply the attribute of interest–if the question is "what is the shape of the cookie?", any caption that describes or alludes to a shape should be discarded. We query LAION using WIMBD [21] and sample 500 captions for each concept. In a few-shot setup, we provide GPT-4o with the caption, concept, and question, and instruct it to reply with "yes" if both conditions are met, otherwise "no". We download the images associated with the relevant captions. We use AMT crowdsourcing to evaluate the quality of filtering and find our F1 is 90.3% over 1,000 evenly-classified captions. We provide more details in Appendix H.

We compute the distributions based on the images and questions using GRADE. We only use the attribute extraction step, as the image generation step is swapped in favor of sampling images from LAION, and the questions and categories were already generated when we compared the diversity of models. The final result is a total of 100 distributions, each with at least 150 images.

Did the model learn its low diversity from the training set? We use the relevant captions to generate images, and then compare the corresponding distributions. We sample 150 captions from each distribution and use each one to generate 20 images, such that each distribution consists of approximately 3,000 images. We then use GRADE to estimate the distributions and compute the mean entropy, as discussed in Section 4.1.

Did the model generalize its behavior? We compare the distributions from LAION to the corresponding distributions from Section 4.1. We note that this time, while the distributions are defined exactly the same, i.e., the concepts, questions, and categories are the same, the prompts used to generate images are different from the captions in LAION.

Results for both experiments are shown in Table 10. The first experiment provides two evidence. (1) the models closely align with their training data, indicated by the close entropy values, strong positive correlation, and low TVD. (2) the entropy values based on images with underspecified captions in LAION are indeed in par with the entropy values in Section 4.2, which indicates it is a factor in the models' low diversity. The second experiment shows that the models generalize over the training

data, as the trend is similar, albeit to a lesser extent. Together, we pinpoint underspecified captions in the training data as a primary source of low diversity.

Caption filtering. For each concept-level distribution, we use GPT-4o to collect captions that mention the concept but do not indicate the answer to the question. Concretely, we use the prompt below. 400

```
In this task, you are provided with a caption associated with an image, a
concept, and a question. You need to find relevant captions that do not indicate
the answer to the question. Your role is two-part. First, determine whether the
caption explicitly mentions the concept as a tangible thing, and not an
accessory or an item related to the concept. Second, determine if that question
can be answered only by reading the caption. If the answer is yes for the first
and no for the second, reply with "yes", otherwise reply with "no".

Here are some examples to guide your understanding:

Caption: teapot, glass teapot, Chinese teapot, herbal teapot, teaware
Concept: teapot
Question: What material is the teapot made of (ceramic, metal, glass, etc.)?
Reasoning: The first part is to determine if teapot is mentioned in the prompt.
It is the first word in the caption, so it is. The second part is to determine
if the question is answerable from the prompt or not. We want to find captions
that are not answerable. Since there are mentions of materials in the caption,
it is answerable and the answer is no.
Answer: no

Caption: My Sweet Angel Book Store Hyatt Book Store Amazon Books eBay Book Book
Store Book Fair Book Exhibition Sell your Book Book Copyright Book Royalty Book
ISBN Book Barcode How to Self Book
Concept: book
Question: Is the book dirty or clean?
Reasoning: The caption mentions items related to a book, but not an actual book.
The answer is no.
Answer: no

Caption: Perfect reading chair, cozy reading chair, nest chair, my favorite
chair, Nest Chair, Cozy Chair, Chair Cushions, Big Chair, Cuddle Chair, Swivel
Chair, Relax Chair, Big Comfy Chair, Chaise Chair
Concept: chair
Question: What color is the chair?
Reasoning: The first part is to identify if the caption mentions a chair. It
does mention a chair, with various adjectives. The second part is to determine
if the question is answerable from the caption. If the question is answerable,
my final answer is no, if it is not answerable, I answer yes. The question asks
about the color of the chair, and there is no mention of a chairs color. The
answer is yes.
Answer: yes

Caption: JIX motorcycle helmet, cross helmet, full helmet, safety helmet
Concept: helmet
Question: Does the helmet have any logos or graphics on it?
Reasoning: The first part is to determine if the caption mentions a helmet. The
caption indeed mentions a variety of helmets. The second part is to determine if
the question can be answered from the caption alone. There is no information
about logos or graphics in the caption, so it is not answerable from the caption
alone. The final answer is yes because the answer to the first is yes and the
second is no.
Answer: yes

Caption: dust bin, garbage container, recycle bin, trash icon
Concept: bin
```

```
Question: What shape is the bin?
Reasoning: The first part is to determine if the caption mentions a bin. The
caption mentions a bin, but it also mentions trash icon. This indicates this is
not an actual bin, but an icon of a bin. The answer is no.
Answer: no

Caption: Cookie Policy - Cookie Law Compliance [MultiLang..
Concept: cookie
Question: What shape is the cookie?
Reasoning: The first part is to determine if the caption mentions a cookie. The
caption mentions cookie policy and cookie law compliance, but not an actual
edible cookie, that has a shape. The answer is no.
Answer: no

Caption: Best Cookie Presses - Cookie Press 150PCS Cookie Press Gun with 16
Review
Concept: cookie
Question: Does the cookie have chocolate chips?
Reasoning: The first part is to determine if the caption mentions a cookie or
something else. The caption is about cookie press and not actual cookie. The
answer is no.
Answer: no
```

H Human Evaluation

Worker selection. Workers were chosen based on their performance records, requiring them to have a minimum of 5,000 approved HITs and an approval rate above 98%. They had to achieve a perfect score on a qualification exam before being granted access to the task. An hourly wage of $15 was provided, ensuring they were fairly compensated for their efforts.

Validating GRADE. To validate the VQA Appendix D.2, we run an AMT task where the worker is provided with a question, concept, image, and categories, and is requested to select the category that best matches the question and image. The UI for this task can be viewed in Fig. 8 with examples in Fig. 9. A sample of cases from our category coverage validation is available in Fig. 10 and Fig. 11.

Validating caption filtering. We run an AMT task to evaluate the caption filtering of GPT-4o from Appendix G. We instruct workers to identify if our conditions are met: (1) the caption mentions the concept itself not something peripheral to it (e.g., a cookie and not cookie cutter); and (2) the caption does not indicate an answer to the question (attribute) of the distribution e.g., "what is the shape of the cookie?". We sample 1,000 captions, 500 from each class (relevant and not relevant). We use three workers per HIT and select the majority decision. We find that GPT-4o scores 85.8% recall and 95.4% precision, making it a reliable automated filtering method. The UI for this task can be viewed in Fig. 12 and a sample of example cases can be viewed in Fig. 13.

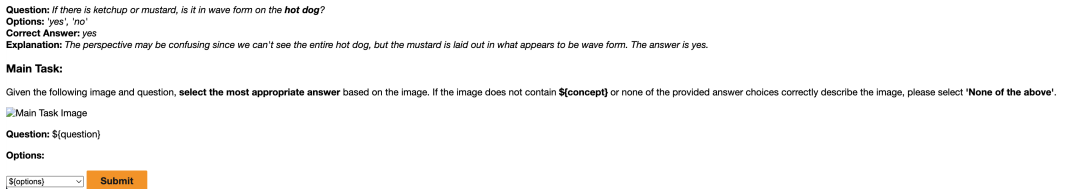


Figure 8: A screenshot of the VQA validation task. Workers are provided a question, concept, image, and a set of categories, including "none of the above" (options here). Their task is to select the option that answers the question.

Instructions: In this task, you will be provided with an image and a question. Your job is to select the correct answer to the question based on the options in the dropdown menu. If no option reasonably fits the question or the object you are asked about is not in the image, select the ``None of the above'' option. Below are examples of how to select an answer. Please use it as a guide for the main task that follows.

Example 1:

Question: *What type of* ***helmet*** *is depicted in the image (e.g., sports, construction, military)?*
Options: *['aviation helmets', 'diving helmets', 'motorcycle helmets', 'firefighter helmets', 'mining helmets', 'engineering helmets', 'construction helmets', 'ceremonial helmets', 'bicycle helmets', 'equestrian helmets', 'military helmets', 'skiing helmets', 'sports helmets']*
Correct Answer: *bicycle helmets*
Explanation: *The image shows a bicycle helmet.*

Example 2:

Question: *Is the* ***umbrella*** *open or closed?*
Options: *'closed', 'open'*
Correct Answer: *open*
Explanation: *The umbrella is open.*

Example 3:

Question: *Is the* ***drawer*** *open or closed?*
Options: *'open', 'closed'*
Correct Answer: *None of the above*
Explanation: *There is no drawer in the image, there is something that looks like a table, but it does not have an inner shelf for item storage.*

Figure 9: 3 out of 10 examples provided to workers as aid to complete their visual question answering task.

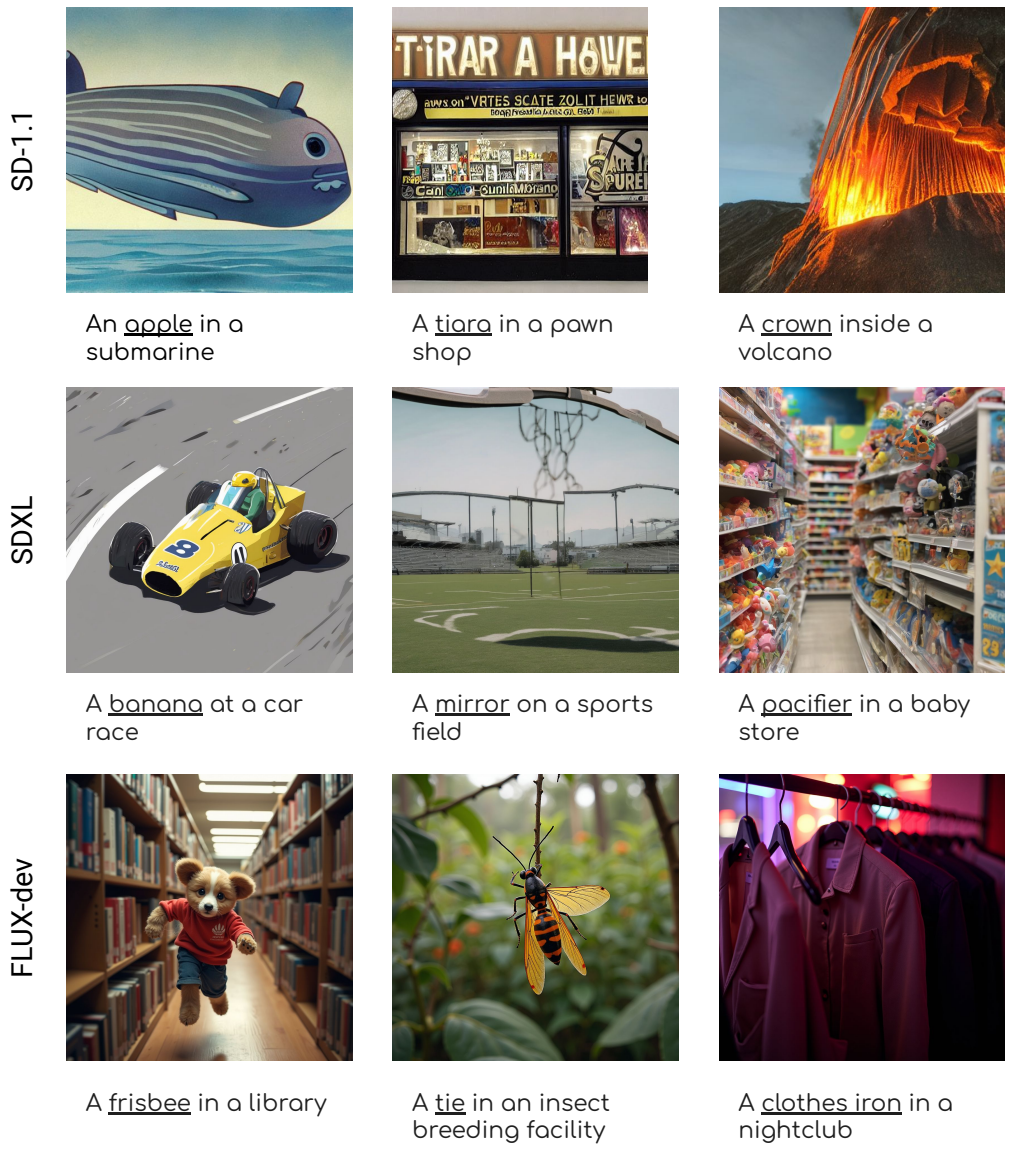


Figure 10: A sample of images marked with "none of the above", as a result of not including the concept (underlined) in the image.

SDXL

Popcorn at a cinema

Q: is the popcorn in a bowl or a bucket?

A_c = {bucket, bowl}

SD-1.1

a toy at a children's playroom

Q: Does the toy appear to be mechanical or electronic?

A_c = {mechanical, electronic}

SDXL

a tie in an office

Q: Is the tie worn with a formal or casual outfit?

A_c= {casual, formal}

FLUX-dev

A person in a city square

Q: Is the person male or female?

A_c = {male, female}

Figure 11: A sample of images marked with "none of the above". The top row exhibits cases where the category is not in $\mathcal{A}_c$. The bottom row exhibits cases where the question cannot be answered just from viewing the image.

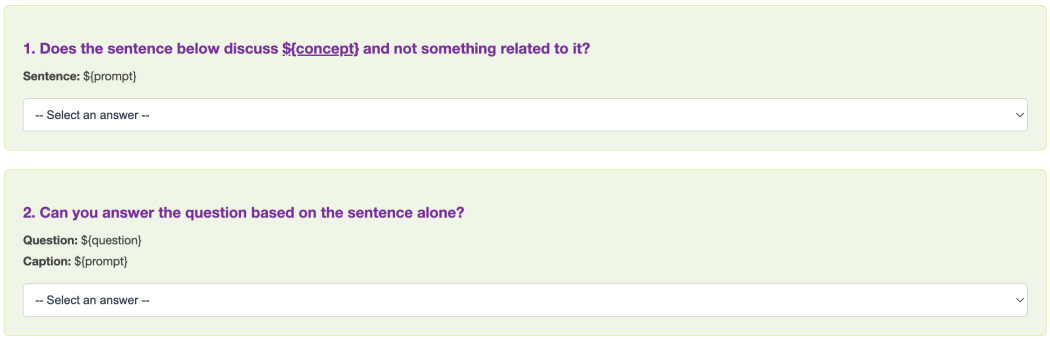


Figure 12: A screenshot of the caption filtering validation task. Workers are provided a caption, two questions, and a concept. Their task is to read the caption and answer the questions.

Instructions:

You will be presented with sentences and questions about them. Your task is to read each sentence carefully and answer two questions:

1. **Does the sentence below discuss ${concept} and not something related to it?**
2. **Can you answer the question based on the sentence alone?**

For each question, select **"Yes"** or **"No"** based on the following guidelines:

- **For Question 1:**
  - Select **"Yes"** if the sentence directly discusses the specified concept and not something related to it.
  - Select **"No"** if the sentence does not discuss the concept directly or discusses something related but not the concept itself.
- **For Question 2:**
  - Select **"Yes"** if you can answer the question based solely on the information provided in the sentence.
  - Select **"No"** if you cannot answer the question based solely on the sentence, or if additional information is required.

Please refer to the examples below for guidance:

> *Sentence:* O'Neal - Q RL Helmet - Bicycle helmet
> *Does the sentence below discuss a helmet?* **Yes**
> *Explanation:* The sentence says it is a bicycle helmet.
> *Question:* What type of helmet is depicted in the image (e.g., sports, construction, military)?
> *Can you answer the question based on the sentence?* **Yes**
> *Explanation:* This is a bicycle helmet, as stated in the sentence.

> *Sentence:* Motorcycle Helmet Motocross Helmet cookie cutter set
> *Does the sentence below discuss a helmet?* **Yes**
> *Explanation:* The helmet is a motorcycle helmet, so we know it's an actual helmet.
> *Question:* What color is the helmet?
> *Can you answer the question based on the sentence?* **No**
> *Explanation:* The sentence doesn't imply the color of the helmet.

> *Sentence:* Photo #2 - Cookie & Cookie Monster
> *Does the sentence below discuss a cookie?* **Yes**
> *Explanation:* The sentence explicitly mentions "Cookie," identifying it as a concept in the sentence.
> *Question:* What shape is the cookie?
> *Can you answer the question based on the sentence?* **No**
> *Explanation:* The sentence does not provide information about the shape of the cookie, only its presence.

Figure 13: 3 out of 10 examples provided to workers as aid to complete their caption filtering task.

I Examples of low diversity

Figure 14: 100 images of “a princess at a children’s party” using FLUX.1-dev. All depictions of princess are of a Caucasian child wearing a pink dress (except two). The background is similar in all of them, often including another princess wearing a blue dress.