GRADE: Quantifying Sample Diversity in Text-to-Image Models

Anonymous ACL submission

Abstract

Text-to-image models are mainly judged by prompt adherence and realism. Yet, prompts are inherently underspecified, leaving ample room to realize them in diverse ways. For instance, "a princess at a children's birthday party" does not detail the princess' outfit, or demographic traits like skin color, age, etc. Yet, models *consistently* depict her as the birthday girl with similar skin tone and a pink dress. In those common, underspecified cases, we would expect models to cover the full spectrum of the concept. Measuring and quantifying this ability has been mostly unexplored. In this work, we present **GRADE**, a natural-language driven method to quantify *sample diversity* across concept-specific attributes. We use GRADE to measure diversity in 12 models and reveal all models collapse to *default behaviors*, a phenomenon where a model consistently generates concepts with the same attributes (e.g., 98% of dresses are pink). Notably, we find that diversity often worsens with stronger prompt adherence and larger models. Lastly, we attribute underspecified captions in the training data for models' low diversity, where omitted attributes correlate with default behaviors.

1 Introduction

Text-to-image (T2I) models have the remarkable ability to generate realistic images from textual descriptions. Yet prompts are inherently *underspecified* (Hutchinson et al., 2022; Rassin et al., 2022), they rarely constrain the model to a specific subtype of a concept. In principle, this should allow the model to sample across the wide range of valid variations. For instance, the phrase "a cookie in a bakery" could plausibly yield cookies of different shapes, colors, and textures. However, in practice, models often collapse onto the same narrow subtype, producing visually similar outputs despite the openness of the prompt. This tendency raises the question: to what extent do current T2I models explore the full spectrum underspecified descriptions afford? Measuring this diversity remains difficult, as the space of valid attributes is effectively unbounded.

Figure 1: GRADE scores for generations and matching search results for three models and concepts. Generated images are substantially uniform with respect to various concepts such as color and size. Search images are included for reference, and show a much greater diversity.

Existing metrics, such as Fréchet Inception Distance (FID; Heusel et al. 2017) and Precision-and-Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) are distribution-based. They measure diversity by comparing representation-level similarities between the outputs of a model to a designated reference set, typically its training set (which is often not diverse either). Because these metrics are representation and distribution based, their ability to isolate specific attributes of a concept is limited, and the diversity score is not easily traced back to a specific factor, making it hard to interpret.

Our desiderata for a diversity metric are for it to capture granular, concept-specific variations, be human-interpretable, and reference-free.

We propose **Gr**anular **A**ttribute **D**iversity **E**valuation (GRADE), a method for measuring sample diversity in T2I models at a granular,

concept-dependent manner, focusing on attributes, such as the *shape* of a *cookie* or the *state* of an *umbrella*. Our approach (illustrated in Figure 2) involves using a large language model (LLM) to generate prompts that elicit diverse outputs from T2I models. These prompts are accompanied by questions that tailor common, specific *attributes*–relevant axes of diversity–for each concept (e.g., "What is the shape of the cookie?" and "Is the umbrella open or close?"). We use a visual question-answering (VQA) model to extract attribute values from images using the questions. Separate to the VQA, we use an LLM to approximate the value set (a set of plausible attribute values) based on the concept and attribute alone. Finally, we map the VQA outputs to values in the attributes set. The result is a categorical distribution over a concept and an attribute. We compute its normalized entropy and use it as our diversity score.

Using GRADE, we determine that no model we test is particularly diverse, with the highest score being 0.64 on a scale from zero to one. For example, one of the state-of-the-art models, FLUX.1-dev, advertised as a highly diverse model, produces strikingly uniform images for "a princess at a children's party", scoring only 0.22 (see Figure 1); here, the princess is consistently white, with a dress and tiara–a phenomenon we call *default behavior*. We explain such scores from captions omitting common attribute values in the training data, resulting in reporting bias, previously explored in societal bias contexts (Seshadri et al., 2023; Luccioni et al., 2023). Our contributions are three-fold:

- **A diversity evaluation method:** We introduce GRADE, a fine-grained and interpretable method for measuring attribute-level diversity of T2I models in underspecified concepts.
- **Comparative diversity analysis:** Using GRADE, we conduct an extensive study comparing the diversity of 12 T2I models across 720K images, revealing that even the most diverse ones achieve low diversity and consistently exhibit default behaviors. Interestingly, our analysis uncovers negative correlation between model size and diversity.
- **Reporting bias in training data:** We demonstrate using LAION and Stable Diffusion (SD) models that underspecified captions in the training data contribute to low diversity of underspecified prompts.

2 Related Work

Most diversity measurements are *distribution-based*: a set of images generated by the evaluated model is compared to a reference set that captures the desired diversity, typically in feature-space, using an image encoder such as Inception v3 (Szegedy et al., 2014; Salimans et al., 2016) or CLIP (Radford et al., 2021).

Perhaps most popular, Fréchet Inception Distance (FID) outputs a score representing both fidelity and diversity and is the standard for evaluating image generating models. However, it has numerous documented issues, like numerical sensitivity, data contamination, and biases (Parmar et al., 2022; Bińkowski et al., 2018; Chong and Forsyth, 2020; Kynkäänniemi et al., 2022; Jayasumana et al., 2024). Precision-and-Recall (Sajjadi et al., 2018) separated fidelity and diversity to two metrics. Additional metrics were proposed (Kynkäänniemi et al., 2019; Naeem et al., 2020; Kim et al., 2023; Alaa et al., 2022), which decouple between different properties and offer more interpretable methods. Crucially, all these methods rely on a set of **diverse** reference images, by comparing the distribution of generated images to the reference set with the desired level of diversity. This can be the model's training data, or an established dataset, like ImageNet (Deng et al., 2009). However, acquiring reference images that faithfully reflect diversity is not straightforward and often requires using an encoder trained on similar data to capture the similarities between the distributions. These requirements make it difficult to reproduce the results of previous work and maintain the integrity of the metrics as they are sensitive to data contamination, which could make them favor models that produce patterns similar their training set, regardless of diversity (Kynkäänniemi et al., 2022).

Perhaps more importantly, these metrics measure the diversity of images as a whole, while we measure diversity over **specific** semantic axes. For example, if we use the above metrics to compare two nearly-identical images of a bottle, with only its color as the difference, the output score would be very high. However, our metric would capture such difference, as we show in Section C.

Similar to GRADE, Vendi Score (Friedman and Dieng, 2022; Pasarkar and Dieng, 2023) is a reference-free metric, defined as the entropy of the eigenvalues of a user-provided similarity metric. However, it is sensitive to the selected metric

and is not granular and natural-language-driven. OpenBias (D'Incà et al., 2024) is an automatic bias detection method that leverages LLMs and VQAs to identify open-sets of biases in T2I models, with a focus on assessing fairness. Although GRADE takes a similar approach, our purpose is inherently different: to quantify and interpret an overlooked aspect in T2I models–the attribute variability of underspecified concepts in prompts.

3 GRADE: Measuring Sample Diversity

3.1 Approach

We seek to quantify the variability of images produced by a T2I model for a given concept c when the prompt underspecifies certain attributes. Concretely, let C be a random variable representing possible *concepts* (*e.g.*, "cookie") and let A be a random variable representing *attributes* (*e.g.*, "shape"). An attribute $A = a$ may take values in a set $\mathcal{V}_c^a$, denoting how a can manifest for concept c (*e.g.*, a cookie's shape could be "round" or "square").

To characterize the probability of observing an attribute value $v \in \mathcal{V}_c^a$ in a generated image, we define the *concept distribution*:

$$P_{V|a,c}(v) \;=\; P\big(V = v \mid A = a,\, C = c\big) \quad (1)$$

Ideally, one would generate *all* possible images of c to empirically determine the frequency of each value v. However, both the conceptual and attribute spaces can be immense, making exhaustive enumeration infeasible. Moreover, identifying which attributes apply to a given concept necessitates world knowledge (for instance, "open or closed" is relevant for an umbrella but not for a cookie).

Instead, we approximate the attribute set $\mathcal{V}_c^a$ with $\tilde{\mathcal{V}}_c^a$ based on language-model-derived world knowledge. We also define a set of *underspecified prompts* $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, each referencing concept c while leaving target attributes unspecified. We then obtain a *multi-prompt distribution*:

$$\tilde{P}_{V|a,c}(v) \;=\; \frac{1}{n}\sum_{i=1}^{n} P\Big(V = v \mid A = a,\, C = c,\, p_i\Big) \quad (2)$$

which reflects, across multiple prompts, how frequently the T2I model generates attribute value v for concept c.

To measure diversity, we compute the normalized entropy of $\tilde{P}_{V|a,c}$:

$$\hat{H}\big(\tilde{P}_{V|a,c}\big) \;=\; \frac{H\big(\tilde{P}_{V|a,c}\big)}{\log_2\big|\tilde{\mathcal{V}}_c^a\big|} \quad (3)$$

where $H(\cdot)$ is the Shannon entropy and $|\tilde{\mathcal{V}}_c^a|$ is the cardinality of the approximate attribute-value set. By definition, $\hat{H}$ ranges from 0 (all images collapse onto a single attribute value) to 1 (the attribute values are evenly distributed). We will refer to $\hat{H}$ simply as the "entropy" for brevity.

Although we focus on *multi-prompt distributions*, one can also examine *single-prompt distributions*, which measure how a single prompt $p \in \mathcal{P}$ distributes across the attribute values in $\tilde{\mathcal{V}}_c^a$. Averaging these metrics across concepts and attributes yields a global measure of a T2I model's diversity.

3.2 Method

Our proposed GRADE pipeline comprises four steps, demonstrated in Fig. 2:

(a) Prompting concepts. We first design two kinds of underspecified prompts for each concept c: *common prompts*, which situate c in familiar or high-frequency scenarios that often appear in web-scale training corpora, and *uncommon prompts*, which deliberately embed c in rare or surprising contexts. Common prompts (*e.g.*, "a cookie during Christmas festivities") may highlight typical attribute-value associations (such as tree-shaped cookies), whereas uncommon prompts (*e.g.*, "a cookie in a volcano crater") test whether the model defaults to certain "usual" attributes even under contextually unusual conditions. This dichotomy reveals whether certain attributes (like shape) are persistently tied to c despite substantial context shifts. Table 2 provides a representative sample of concepts and prompts.

(b) Deriving attributes. Next, we identify relevant attributes for each concept, such as "color", "shape", or "state" (open/closed). To identify them, we query an LLM with just concept c and request it to produce a list of questions about visual attributes (*e.g.*, "What is the *shape* of the cookie?"). In a separate query, we use the LLM to derive a set of plausible answers (*e.g.*, "round", "square") based on each generated prompt from the previous step (*e.g.*, "tree-shaped" is a plausible answer if the prompt describes a cookie during Christmas festivities). Then, we merge all answers into $\tilde{\mathcal{V}}_c^a$ by unifying semantically similar terms (e.g., "round" and "circular" would be unified). This step ensures we capture domain-appropriate attributes for each concept and avoid missing frequently occurring variations. Table 1 provides a representative sample of the generated concepts, attributes, and attribute values, illustrating how $\tilde{\mathcal{V}}_c^a$ is constructed.

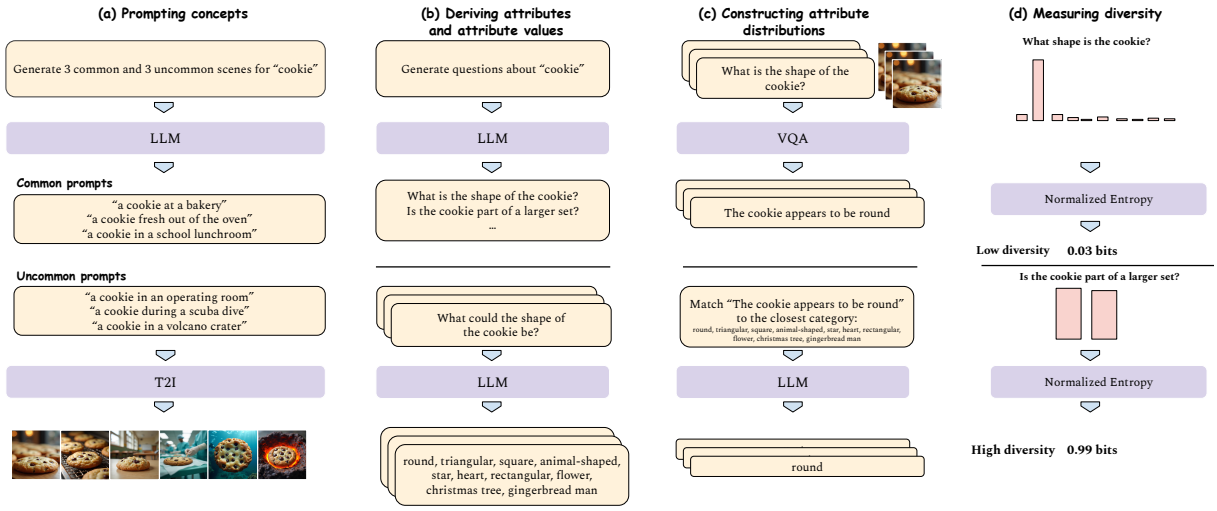


Figure 2: **Workflow of GRADE using "cookie" as input.** (a) Generate prompts that mention "cookie" without specifying its attributes, and use them to generate images. (b) Formulate attribute-related questions and extract responses from the images using a VQA model. (c) Produce attribute values and map the responses to these values. (d) Quantify the diversity of the resulting attribute distributions.

(c) Constructing attribute distributions. For each generated image, we use a VQA to answer the attribute question. An LLM then maps this free-form answer to one of the entries in $\tilde{\mathcal{V}}_c^a$. If no valid match is found (*e.g.*, because the image fails to depict concept c or the answer is not included in the attribute value set), the response is mapped to "none of the above" and is discarded. We then tally remaining attribute values. Repeating this for all prompts in $\mathcal{P}$ yields our approximated multi-prompt distribution $\tilde{P}_{V|a,c}(v)$.

(d) Measuring diversity. We compute the normalized entropy of each distribution to obtain their diversity scores. To derive a score that reflects the diversity of all the measured distributions, we simply aggregate the scores. Higher entropy indicates that a model spreads its generations more uniformly across $\tilde{\mathcal{V}}_c^a$, whereas lower entropy suggests mode collapse toward a particular attribute value.

Implementation details. In step (a), we generate $three$ common and $three$ uncommon prompts per concept, each yielding 100 images for a T2I model. In step (b), we typically have $four$ attributes per concept. Our pipeline uses GPT-4o (OpenAI et al., 2024) (`gpt-4o-2024-08-06`) with temperature 0 and a max token limit of 1,000. Full prompt details are provided in Appendix I.

The cost of estimating a multi-prompt distribution is approximately $0.75, and a single prompt distribution is $0.12, using batch inference. In our experience, wait time is several minutes. Images were generated using an A100-80GB.

Concept	Attribute	Attribute Values
Teapot	Shape	rectangular, spherical, oval, round, square, cylindrical
Person	Singular	yes, no
Suitcase	Vintage	yes, no
Bear	Species	polar, black, sloth, grizzly, sun, panda

Table 1: **Sample of concepts, attributes, and attribute values.** Each concept-attribute pair is a multi-prompt distribution. See Section B. for a larger sample.

4 Comparing Diversity of Models

We use GRADE to estimate the diversity of popular T2I models. We begin with an overview of our setup and then present the results.

Data and distributions overview. For each model, we estimate distributions over 100 common concepts such as "cookie" and "suitcase" and attributes such as "shape" and "color". Each concept is linked to four questions on average. In total, there are 405 multi-prompt distributions and 2,430 single prompt distributions, consisting a total of 60,000 images per model.

T2I models. We use 12 models from three families. **IF-DeepFloyd** (at StabilityAI, 2023) includes DeepFloyd-M, DeepFloyd-L, and DeepFloyd-XL. **Stable Diffusion** (Rombach et al., 2022; Podell et al., 2023; Sauer et al., 2023; Luo et al., 2023; Esser et al., 2024) includes SD-1.1, SD-1.4, SD-2.1, SDXL, SDXL-Turbo, SDXL-LCM, and SD-3 (2B). Finally, **FLUX** (Black Forest Labs, 2024a,b) includes FLUX.1-schnell and FLUX.1-dev. All

Concept	Common Prompts			Uncommon Prompts		
Teapot	a teapot in a tea shop.	a teapot in a home kitchen.	a teapot in a British museum.	a teapot at a sports stadium.	a teapot in a space station.	a teapot in a climbing gym.
Person	a person in a city square.	a person in a restaurant.	a person in a concert hall.	a person in an empty desert.	a person on a volcano.	a person underwater.
Suitcase	a suitcase in an airport.	a suitcase at a hotel lobby.	a suitcase at a travel goods store.	a suitcase at a swimming pool.	a suitcase at a construction site.	a suitcase at a forest campsite.
Bear	a bear in a wildlife reserve.	a bear in a zoo.	a bear in a forest.	a bear in a bank.	a bear in a museum.	a bear in a shopping mall.

Table 2: **Sample concepts with common and uncommon prompts.** A larger set is provided in Section B.

	GRADE Score ↑	
Model	**Multi-prompt**	**Single-prompt**
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

Table 3: **GRADE score in multi- and single-prompt distributions.** Values close to 1 indicate highly diverse behavior (uniform) while values close to 0 indicate repetitive generations. The *most* diverse models are in bold. All scores have standard error below 0.01. We omit standard deviation since the underlying entropy distributions are bimodal (Figure 15), making it uninformative.

models were used with the default `Diffusers` library (von Platen et al., 2023) settings.

4.1 Results

All models have low diversity scores. Table 3 presents mean GRADE scores of models across multi- and single-prompt distributions. Permutation tests confirm the observed differences are statistically significant (see Section D.2). The average diversity across all models over multi-prompt distributions is 0.57 and 0.44 over single-prompt distributions, indicating low diversity. Figure 3 illustrates differences in diversity between models, with additional examples in Section A.

Relation of diversity to model size. The trend in Figure 4 indicates an *inverse-scaling law* (McKenzie et al., 2023) between model size and diversity. Specifically, we find that the Pearson and Spearman correlations between diversity and model size are $r = -0.7$ ($p = 0.011$) and $\rho = -0.84$ ($p = 0.001$) respectively. However, given the small sample size of 12 models, and potential confounding factors, such as different data and architectures, we do not make any causal claims and these findings should be interpreted with caution. Furthermore, in addition to our claims in Section 6 (that underspecified captions cause low diversity), Figure 4 shows that the more images by a model are mapped to "none of the above"[1] (i.e., prompt adherence *decreases*), the more diverse it is. Pearson $r = 0.8$ ($p = 0.02$) and Spearman $\rho = 0.94$ ($p < 0.001$) correlations reinforce this, suggesting the possibility that improving the ability of models to generate images that match the prompt is at the cost of sample diversity, similar to fidelity-diversity tradeoffs shown before (Dhariwal and Nichol, 2021; Kynkäänniemi et al., 2019).

Default behaviors. Across all concepts we test, we observe a highly recurring phenomenon we call *default behaviors*. These occur when a concept is consistently ($\geq 80\%$ of instances) realized with the same attribute value, despite the attribute being underspecified in the prompt. Notably, default behaviors are present in distributions scored low by GRADE. To quantify its prevalence, we measure the rate at which models default to the same attribute value across repeated generations, both within a single prompt and across multiple prompts. Section E includes aggregate results for models.

	% of Default Behavior ↓	
Model	**Multi-prompt**	**Single-prompt**
DeepFloyd-M	83	92
DeepFloyd-L	81	92
DeepFloyd-XL	80	92
SD-1.1	78	87
SD-1.4	82	87
SD-2.1	76	89
SDXL	81	90
SDXL-Turbo	86	95
SDXL-LCM	82	92
SD-3 (2B)	88	95
FLUX.1-schnell	**90**	**97**
FLUX.1-dev	88	96

Table 4: **Percentage of at least one default behavior.** Lower values indicate higher diversity. Almost all concepts are associated with at least one default behavior in single prompt distributions, with a similar trend in multi-prompt distributions. The model with the *most* default behaviors is in bold.

Results. We find that even the distributions from the **most** diverse model (SD-1.1) collapse **87%** of

[1] In Section 5 we show that 80% of unanswerable images do not depict the concept mentioned in the prompt.

Figure 3: **Images generated with the prompt "a princess at a children's party" show differences in model diversity.** From top to bottom, SD-1.4 (most diverse), SDXL, and FLUX.1-dev (least diverse). Although none are highly diverse, there is a marked difference between them. Specifically, diversity is reduced in attributes such as the ethnicities of depicted people, colors of dresses, and overall backgrounds.

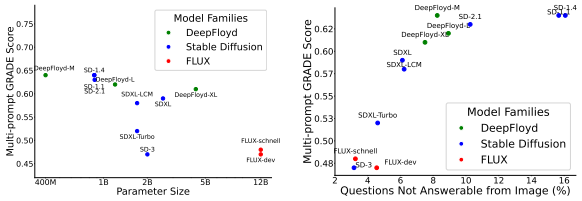


Figure 4: **(a) GRADE score in multi-prompt setting plotted against the denoiser's parameter size.** To a degree, diversity deteriorates as parameter size grows. This effect is most apparent within every model family. **(b) GRADE score in multi-prompt setting plotted against percentage of answers mapped to "none of the above".** In Section 5 we show 80% of which account for missing concepts in images. Low "none of the above" values correspond to *high* prompt adherence. The plot suggests an adherence-diversity tradeoff.

the time in a single-prompt setting, and **76%** in a multi-prompt setting. These patterns exemplify the entropy trends in Figure 4. Complete results with further analyses are provided in Section E. To summarize, models tend to generate a concept with the same attribute value when sampling it many times over the same prompt, and even when varying the prompts (but maintaining the concept underspecified), diversity only slightly improves.

5 Validating GRADE

Here, we describe the validation process, which is provided in full detail in Section G. We break down the validation process to two parts. First, we manually verify that the data was generated correctly. For prompts, we confirm all are underspecified (contain a concept without attributes), and to make sure the generated prompt is common or uncommon, we extract the nouns in the prompt and check their count in LAION-5B. Indeed, we find all prompts are generated correctly, and on average, the co-occurrence of nouns in common and uncommon prompts is 30,655 and 956, respectively. We also verify that each question can be answered solely by viewing an image of the concept. For the answer sets, we verify that no two answers semantically overlap (e.g., "circle" vs. "round"). Indeed, we find that all data was generated according to our definitions in Section 3.1.

Next, we validate the quality of the VQA component. We evaluate its ability to correctly answer the question, and its ability to identify out-of-scope questions, where either the model failed to adhere to the prompt, or that the answer set is insufficient to answer the question. We use 71 qualified AMT workers to answer the same questions over 2,800 sampled images and find an agreement of 91.8% between the VQA and a 3-way human majority.

5.1 Comparing GRADE to Previous Metrics

We compare GRADE with two widely used metrics: FID (Szegedy et al., 2014) and Recall (Sajjadi et al., 2018), both work by comparing feature-level

distributions, typically between a model's training set and its generations. To facilitate this comparison, we modify GRADE to operate as a reference-based metric, and replace its entropy term with Total Variation Distance (TVD).

Since LAION is open-source and was used to train SD-1.1, SD-1.4, and SD-2.1; LAION-2B for the first two and LAION-5B for the latter–we sample images from the training set used for each model and compare them to images generated by these models. Specifically, we sample 50 multi-prompt distributions from Section 4 (only attributes and their values). Then, for each reference set, we sample 115 image-caption pairs from the LAION-2B and LAION-5B datasets, using WIMBD (Elazar et al., 2024), where the image depicts the concept and the caption mentions the concept but is otherwise underspecified, in accordance with our approach (Section 3.1). This results in 50 reference distributions for each dataset, each consisting of 115 images. To create matching model distributions, we generate one image using each caption.

We observe that FID is nearly uncorrelated with TVD ($\approx 4\%$), while Recall is negatively correlated ($\approx -18\%$). This divergence possibly arises because FID and Recall summarize distributions in feature space (e.g., via CLIP (Radford et al., 2021)), which may overlook fine-grained attribute variations (e.g., different shapes for a concept like "cookie"). In contrast, GRADE explicitly models attribute values grounded in human-understandable questions (e.g., "Is the cookie round or square?"), thus capturing concept-level diversity that global feature statistics fail to discern.

These findings coupled with the human-based validation, confirm GRADE effectively measures semantic-level variation missed by FID or Recall. Further detail, including analogous analyses using Inception v3 (Szegedy et al., 2014) as a feature extractor and visual examples, reinforce these conclusions (see Section C).

6 Low Diversity Begins in Training Data

In Section 4, we showed that T2I models often exhibit limited diversity when faced with underspecified prompts. We posit that this phenomenon stems from the nature of the training data: whenever a concept is mentioned without an explicit attribute value (e.g., "banana" rather than "yellow banana"), the accompanying images in the dataset tend to be dominated by a small set of attribute values. We observe anecdotal evidence for this in LAION: sampling 100 image-caption pairs that mention a concept without specifying its attribute typically yields images that share an implicit, most common attribute value (e.g., bananas tend to be yellow). This is closely related to the linguistic phenomenon of *reporting bias* (Gordon and Van Durme, 2013), where attributes deemed "obvious" or "typical" are not explicitly mentioned in captions.

Formally, each training example in a T2I dataset is a caption-image pair. When the caption includes a concept but omits an attribute (e.g., "banana"), we hypothesize that the distribution of actual images is heavily skewed toward a small subset of attribute values (most bananas in LAION are indeed yellow). As a result, the model learns to replicate this limited distribution whenever it encounters an underspecified prompt. In what follows, we verify this by comparing the distributions of (i) real images from LAION where captions omit an attribute, and (ii) generated images produced by the same underspecified captions or by similar prompts.

6.1 Experimental Setup and Metrics

To examine this, we use GRADE to measure diversity across multiple prompts (i.e., the *multi-prompt distribution*). In particular, we measure:

- **Training data distribution:** We select underspecified captions from LAION (e.g., "cookie" but not "cookie cutter," and with no mention or implication of a specific attribute). We refer to these as *filtered captions*.

- **Model-generated distribution:** We use the same underspecified captions (and also additional unseen prompts) as inputs to a T2I model and generate multiple images per prompt. We then measure the distribution of attribute values across these generated images.

We compare these distributions using three statistics: (1) *entropy*, which captures overall diversity; (2) Pearson correlation coefficient (PCC), which captures the extent to which the attribute-value frequencies align between the training data and the generated images; and (3) TVD, which measures the dissimilarity between the two distributions.

Replication of training data diversity. First, we test whether T2I models reproduce the diversity observed in underspecified caption-image pairs from their own training data. For each model, we

select 50 triplets of concepts, attributes, and attribute values. We filter LAION captions that mention the concept as an object but do not specify or imply the attribute (e.g., "a cookie on a table" as opposed to "a classic chocolate chip cookie," which implies it is round). We then:

1. Collect up to 150 such *filtered captions* per concept from LAION (technical details are presented in Section F).
2. Compute GRADE on the images associated with these captions.
3. Generate 20 images per filtered caption using a T2I model, resulting in 3,000 generated images per concept.
4. Compute GRADE on these generated images to obtain their distribution of attribute values.
5. Compare the real (LAION) and generated distributions via entropy, PCC, and TVD.

Generalizing to new underspecified prompts. We next explore whether the model's tendency to mirror training data extends to prompts that are not sampled from LAION. Specifically, we compare the multi-prompt distributions obtained in Section 4 with the corresponding distributions from LAION for the same concept-attribute pairs. This comparison reveals whether the model continues to replicate the underspecified distributions it observed in the training set.

6.2 Results

Table 5 summarizes the outcomes. LAION itself exhibits moderate diversity for our selected concepts, reflected by dataset entropy values of 0.64 in LAION-2B and 0.65 in LAION-5B. When prompted with the *exact filtered captions* from LAION, models achieve a similar range of entropy (0.62–0.68). The correlation between model outputs and LAION images is high (PCC of 0.73–0.88), and the TVD remains low (0.10–0.13). These observations imply that T2I models replicate the underspecified distributions seen in their own training data. When the same models are provided with *new*, underspecified prompts ("Generated" in Table 5), the alignment with LAION images diminishes slightly. The PCC drops (0.61–0.72 vs. 0.73–0.88), and the TVD increases marginally (0.17–0.18 vs. 0.10–0.13). Yet, the overall trend remains the same: the generated multi-prompt distributions still resemble those in LAION for the given concept-attribute pairs.

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

Table 5: Similarities between model outputs and their training sets. Entropy, PCC, and TVD show models have comparable diversity to training data.

These results strongly support our core hypothesis: when concept-attribute pairs are left unspecified in captions, most images in the training data depict a single implicit, most common attribute value. Consequently, T2I models learn to replicate this bias. Unless the user explicitly overrides it with a specific attribute, the model reproduces the distribution it has observed most frequently in training, resulting in a systematic lack of diversity.

7 Conclusion

The ability of T2I models to generate diverse attributes based on input prompts was not measured in previous works. In this work, we present GRADE, a method to measure this ability for underspecified concepts in the prompt. We use GRADE to show that 12 popular models collapse to default behaviors, a phenomenon where models depict a concept with the same attribute values, despite the prompt not mentioning the descriptors. We show that larger and better models become even less diverse. This points a blind spot in state-of-the-art models, and existing metrics, since sample diversity is not systematically measured, it is not improved, but deteriorates. Finally, we demonstrated that one culprit to low diversity is reporting bias in the training set, where captions often omit meaningful attributes of a concept, and thus result in biases in the model. We encourage researchers to adopt our work to benchmark sample diversity in T2I models and future refinements, such as improving training data quality or designing diversity-driven model objectives. Exploring multi-attribute relationships, combining GRADE with other evaluation measures, and investigating training interventions are promising directions for future work.

Ethical Considerations

Bla bla

Limitations

While GRADE provides a fine-grained view of sample diversity, it has two limitations. First, as with any metric focused on a specific set of concepts and attributes, its scores depend heavily on which attributes are measured; attributes not included in the evaluation remain unassessed. Second, it relies on external LLM and VQA components, introducing potential biases and inaccuracies from these models into both the attribute-suggestion process and the final diversity score. Despite these limitations, we believe that GRADE represents a step toward more interpretable, fine-grained diversity assessments in T2I models.

References

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. 2022. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR.

DeepFloyd Lab at StabilityAI. 2023. 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. Retrieved on 2023-11-08.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Black Forest Labs. 2024a. FLUX.1-dev Model Documentation. https://huggingface.co/black-forest-labs/FLUX.1-schnell. Accessed: Aug 24 2024.

Black Forest Labs. 2024b. FLUX.1-dev Model Documentation. https://huggingface.co/black-forest-labs/FLUX.1-dev. Accessed: Aug 24 2024.

Stefano Bonnini, Getnet Melak Assegie, Trzcinska Kamila, and 1 others. 2024. Review about the permutation approach in hypothesis testing. *Mathematics*, 12:2617–1.

Min Jin Chong and David Forsyth. 2020. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079.

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

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. 2024. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235.

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. 2024. What's in my big data? *Preprint*, arXiv:2310.20707.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*.

Dan Friedman and Adji Bousso Dieng. 2022. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*.

Jonathan Gordon and Benjamin Van Durme. 2013. 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. Association for Computing Machinery.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv

Kumar. 2024. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315.

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

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2022. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*.

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

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*.

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

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, and 1 others. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR.

OpenAI. 2024. Introducing structured outputs in the api. Accessed: 2024-09-17.

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, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420.

Amey P Pasarkar and Adji Bousso Dieng. 2023. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. *arXiv preprint arXiv:2310.12952*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. 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.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

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

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. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. 2023. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*.

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

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. 2023. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

A Qualitative Examples of Diversity

Figure 5: **Difference in diversity between models.** Images generated using the prompt "a bag on a cliffside". Each row corresponds to a model, top-down: SD-1.4 (most diverse), SDXL, and FLUX.1-dev (least diverse). While no model exhibits high diversity, there is a marked difference between SD-1.4 and FLUX.1-dev, with SDXL between them. Specifically, diversity is reduced in attributes such as color and placement of the bags, as well as the background.

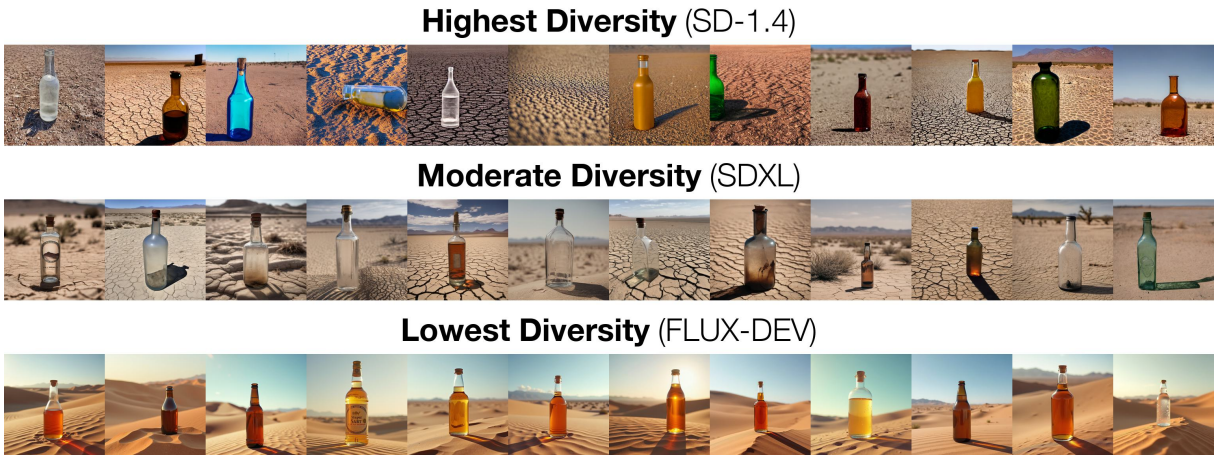


Figure 6: **Difference in diversity between models.** Images generated using the prompt "a bottle in a desert". Each row corresponds to a model, top-down: SD-1.4 (most diverse), SDXL, and FLUX.1-dev (least diverse). While no model exhibits high diversity, there is a marked difference between SD-1.4 and FLUX.1-dev, with SDXL between them. Here, the lack of diversity is most pronounced in the color of the bottle or its liquid. While SD-1.4 depicts relatively varied bottles, SDXL depicts transparent ones, while FLUX.1-dev depicts almost exclusively orange-like bottles.

Figure 7: **Illustration of GRADE score.** Displayed are 24 of the 100 images generated by FLUX.1-dev using the prompt "A car in a car dealership". The accompanying histogram and the subsequent entropy plot both represent the 100 sample. The GRADE score is 0.78, indicating the color of the cars is relatively diverse.

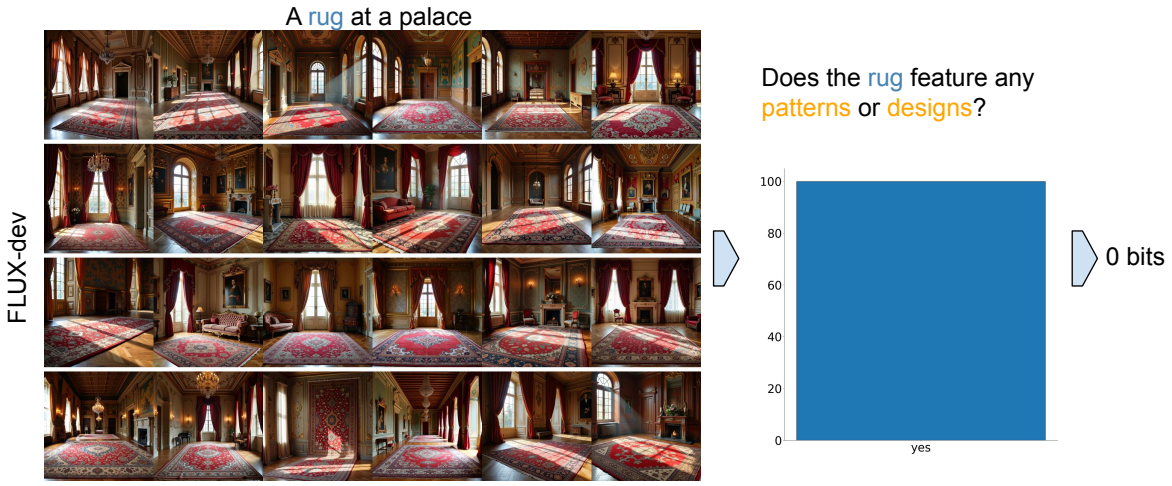


Figure 8: **Illustration of GRADE score.** Displayed are 24 of the 100 images generated by FLUX.1-dev using the prompt "A rug at a palace". The accompanying histogram and the subsequent entropy plot both represent the 100 sample. The GRADE score is 0, indicating the rugs are consistently patterned.

Figure 9: **Illustration of GRADE score.** Displayed are 24 of the 100 images generated by FLUX.1-dev using the prompt "A neon sign at a retro diner". The accompanying histogram and the subsequent entropy plot both represent the 100 sample. The GRADE score is 0.99, indicating the location of the signs is uniform.

B Extended Data Overview

Concept	Attribute	Attribute Values
Bin	What shape is the bin?	circular, octagonal, square, cylindrical, triangular, rectangular, round, oval, hexagonal
	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
	Does the bin have a lid?	yes, no
Person	Is the person male or female?	male, female
	Does the image show the person from up-close?	yes, no
Suitcase	Is the suitcase open or closed?	open, closed
	Is the suitcase soft-shell or hard-shell?	soft-shell suitcase, hard-shell suitcase
Cake	Does the cake have multiple tiers?	yes, no
	Is the cake eaten?	yes, no
	What flavor is the cake?	tiramisu, cheesecake, carrot, chocolate, strawberry, vanilla
Pool	Is there anyone swimming in the pool?	yes, no
	What color is the water in the pool?	reflective like a mirror, black, clear, green, blue, brown
Teapot	What shape is the teapot?	rectangular, spherical, oval, round, square, cylindrical
Bear	What species of bear is depicted in the image?	polar bear, black bear, sloth bear, grizzly bear, sun bear, panda bear

Table 6: **Sample of concepts, attributes, and attribute values.** Each concept-attribute pair is a multi-prompt distribution.

Concept	Common Prompts	Uncommon Prompts
Teapot	a teapot in a tea shop.	a teapot at a sports stadium.
	a teapot in a home kitchen.	a teapot in a space station.
	a teapot in a British museum.	a teapot in a climbing gym.
Person	a person in a city square.	a person in an empty desert.
	a person in a restaurant.	a person on a volcano.
	a person in a concert hall.	a person underwater.
Suitcase	a suitcase in an airport.	a suitcase at a swimming pool.
	a suitcase at a hotel lobby.	a suitcase at a construction site.
	a suitcase at a travel goods store.	a suitcase at a forest camping site.
Bear	a bear in a wildlife reserve.	a bear in a bank.
	a bear in a zoo.	a bear in a museum.
	a bear in a forest.	a bear in a shopping mall.

Table 7: **Sample of concepts with common and uncommon prompts.**

C Comparing GRADE to Previous Metrics

Model	Dataset	FID-R	FID-P	R-P	FID-TVD	R-TVD	P-TVD
SD-1.1	LAION-2B	0.14	−0.15	0	0.12	−0.15	0
SD-1.4	LAION-2B	0.19	−0.40	0	0	−0.20	−0.15
SD-2.1	LAION-5B	−0.21	−0.48	0	0	−0.19	0.15

Table 8: **PCC between GRADE and traditional metrics paired with CLIP.** FID, Recall (R), and Precision (P) show low to moderate degrees of correlation among each other, while the TVD based on the distributions from GRADE exhibits weak correlations with all of them. This indicates the distributions estimated by GRADE capture diversity existing metrics do not.

Model	Dataset	TVD	FID	Recall	Precision
SD-1.1	LAION-2B	0.15	290	0.12	0.88
SD-1.4	LAION-2B	0.15	276	0.15	0.92
SD-2.1	LAION-5B	0.16	290	0.12	0.94

Table 9: **Evaluation results with traditional metrics paired with CLIP.** Each value in the table is the mean of the metric over the 50 pairs of multi-prompt distributions.

Model	Dataset	FID-R	FID-P	R-P	FID-TVD	R-TVD	P-TVD
SD-1.1	LAION-2B	−0.41	0.23	−0.34	0.14	0.04	0
SD-1.4	LAION-2B	−0.48	0.14	−0.22	0.18	−0.10	0.14
SD-2.1	LAION-5B	−0.12	−0.52	0	−0.16	−0.15	0.13

Table 10: **PCC between GRADE and traditional metrics paired with Inception v3.** FID, Recall (R), and Precision (P) show low to moderate degrees of correlation among each other, while the TVD based on the distributions from GRADE exhibits weak correlations with all of them. This indicates the distributions estimated by GRADE capture diversity existing metrics do not.

Extended results for the comparison between GRADE and traditional metrics described in Section 5.1. Results using CLIP for feature extraction can be viewed in Section C and Section C. Results using Inception v3 (Szegedy et al., 2014) (ImageNet features (Deng et al., 2009)) are in Section C and Section C. Below we detail the process of collecting the image sets and comparing between them.

Reference and generated images. Since LAION is opensource and was used to train SD-1.1, SD-1.4, and SD-2.1; LAION-2B for the first two and LAION-5B for the latter–we sample images from it and compare them to images generated by the models. Specifically, we sample 50 of the 405 multi-prompt distributions (that is, only the concept, attribute, and attribute values, not the prompts and images) in Section 4. Next, we sample 115 image and caption pairs using WIMBD, where the image depicts the concept and the caption mentions the concept but not the attribute, in accordance with our approach (Section 3.1). We end up with 50 reference distributions, each consisting of 115 images. To get our generated images, we generate one image for each caption, to maintain equal proportion between the distributions. For example, if an image in LAION is linked to the caption "Unicorn Cookie", its corresponding distribution will contain an image that was generated using that caption as a prompt.

Details of metrics. Using the 50 pairs of distributions, we can compare GRADE to the metrics. Since entropy is not a reference-based metric, we change it in favor of Total Variation Distance (TVD) and use it on top of the distributions estimated by GRADE. We compute FID and Recall, using features from the open-clip implementation (the `ViT-H/14` variant) (Ilharco et al., 2021; Radford et al., 2021), trained on

Model	Dataset	TVD_G	FID	Recall	Precision
SD-1.1	LAION-2B	0.15	19.67	0.35	0.75
SD-1.4	LAION-2B	0.15	15.0	0.45	0.74
SD-2.1	LAION-5B	0.16	18.67	0.49	0.83

Table 11: **Evaluation results with existing metrics using Inception v3.** Each value in the table is the mean of the metric over the 50 pairs of distributions.

LAION-2B. Recall was computed with $k = 3$. We run the same experiment using Inception v3 features with 64 dimensions.

C.1 Qualitative Metric Comparison Examples

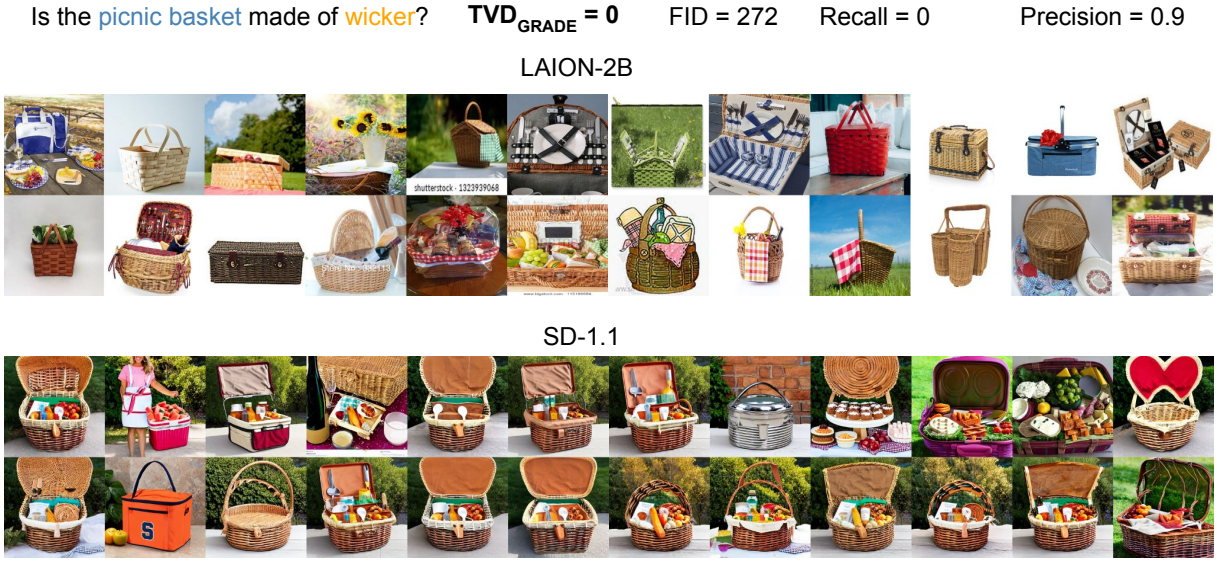


Figure 10: **Comparison between GRADE, FID, and Recall, using CLIP features.** The metrics are compared over the "wicker" attribute of the concept "picnic basket". TVD_GRADE reports very high similarity (lower TVD is better) between the sets of images, which is indeed shown in the images (almost all picnic baskets are made of wicker). In contrast, Recall and FID report very low similarity scores.

Figure 11: **Comparison between GRADE, FID, and Recall, using CLIP features.** The metrics are compared over the "visible stains or damage" attribute of the "tablecloth" concept. $\text{TVD}_{\text{GRADE}}$ reports very high similarity (lower TVD is better) between the sets of images, which is indeed shown in the images (the tablecloth is rarely damaged in either set). In contrast, Recall and FID report very low scores.

D Extended Diversity Comparisons between T2I Models

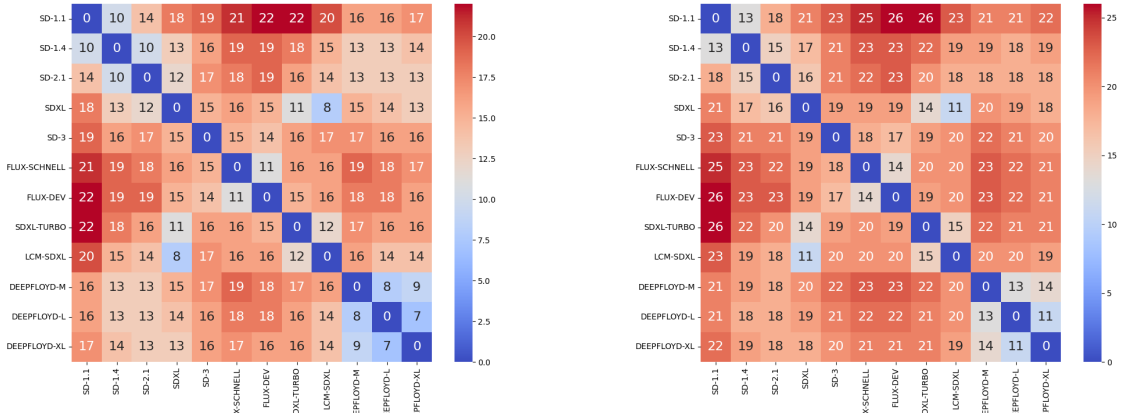


Figure 12: **The mean total variation distance (TVD) between all pairs of models over (a) multi-prompt distributions and (b) single prompt distributions.** For readability, both figures show TVD in a range between 0 and 100 instead of 0 to 1.

Backbone	Models	Mean TVD
SD-1.1	SD-1.1, SD-1.4, SD-2.1	11
SDXL	SDXL, SDXL-LCM, SDXL	10
FLUX	FLUX.1-schnell, FLUX.1-dev	11
DeepFloyd	DeepFloyd-M, DeepFloyd-L, DeepFloyd-XL	8

Table 12: **Backbones, their associated models, and the mean TVD of models with a shared backbone.**

Similarity in diversity across distributions. We investigate the similarity in diversity across models we find in Section 4.1. We modify GRADE to use Total Variation Distance (TVD) instead of entropy to facilitate comparisons between corresponding distributions in the attribute value level. For example, the difference between the frequency of "blue" in the multi-prompt distribution of the concept *tie* and attribute *color*. Results for both multi and single prompt distributions are shown in Figure 12. The results are in line with our other findings: all models have similar distributions, with the maximum TVD for multi-prompt distributions being 0.22 and for single prompt distributions 0.26, with these numbers being the result of a comparison between the least and most diverse models (i.e., SD-1.1 and FLUX.1-dev). Moreover, models with similar backbone have smaller TVDs. The groups and the mean TVDs are shown in Table 12.

D.1 Additional Analysis on Model Size

We further investigate the relationship between model size and diversity, and prompt adherence and diversity. Figure 13 shows that as the denoisers' parameter size increases, the GRADE scores in both the multi and single prompt distributions decrease. This suggests that larger models produce less diverse outputs, indicating an inverse-scaling law (McKenzie et al., 2023). The negative correlation is supported by significant Pearson and Spearman correlation coefficients at both the concept level (Pearson $r = -0.701$, $p = 0.011$; Spearman $\rho = -0.842$, $p = 0.001$) and the prompt level (Pearson $r = -0.666$, $p = 0.018$; Spearman $\rho = -0.804$, $p = 0.002$).

Figure 14 illustrates negative correlation between diversity and prompt adherence. As the percentage of unanswerable images ("none of the above") increases i.e., prompt adherence *decreases*, the diversity measured by entropy increases. This is quantified by strong positive Pearson and Spearman correlations at both the concept level (Pearson $r = 0.802$, $p = 0.002$; Spearman $\rho = 0.938$, ($p < 0.001$) and the prompt

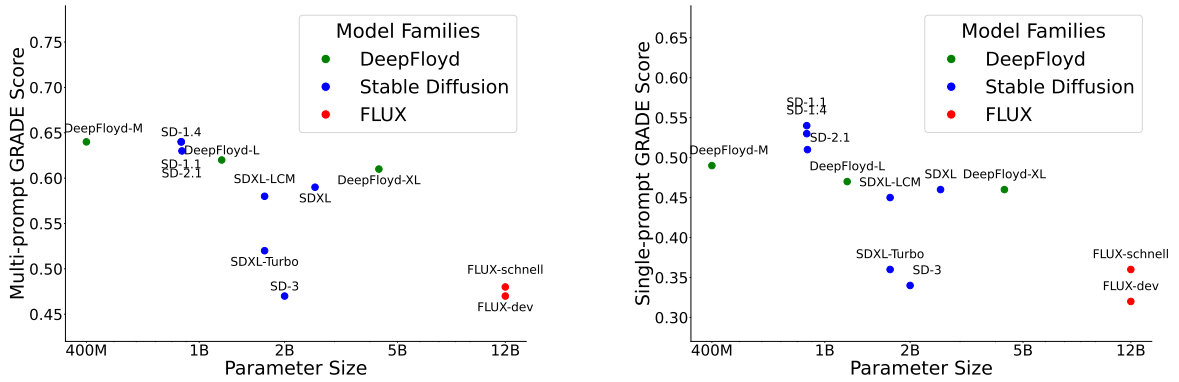


Figure 13: **(a) GRADE score in multi-prompt setting plotted against the denoiser's parameter size. (b) GRADE score in single-prompt setting plotted against the denoiser's parameter size.** To a degree, diversity deteriorates in tandem with parameter size. This phenomenon is most apparent within every model family.

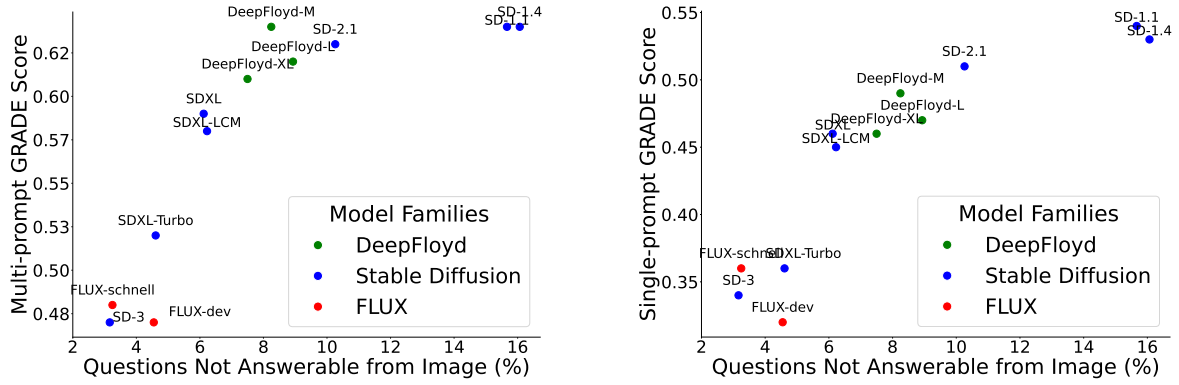


Figure 14: **(a) GRADE score in multi-prompt setting plotted against the % of "none of the above". (b) GRADE score in single-prompt setting plotted against the % of "none of the above".** In Section 5 we show 80% of which account for missing concepts in the image. The plots show negative correlation between diversity and prompt adherence, which indicates there is a tradeoff.

level (Pearson $r = 0.871$, ($p < 0.001$); Spearman $\rho = 0.947$, ($p < 0.001$). This indicates a trade-off between diversity and prompt adherence: models that generate more diverse outputs tend to adhere less strictly to the prompts.

D.2 Statistical Significance of GRADE scores

To confirm our results are statistically significant, we perform a two-tailed permutation test between every unique pair of models for both distribution types (single-prompt and multi-prompt). This test is common when the data comes from a complex distribution (Bonnini et al., 2024), in our case, the distribution of GRADE scores of each model. We demonstrate that the difference between the vast majority of models is statistically significant in both cases.

Concretely, there are 66 unique model pairs. For each pair, we compute a two-tailed permutation test with the null hypothesis H_0 that the GRADE scores of the two models are the same. We perform $N = 100{,}000$ permutations, where the p-value is defined as:

$$p = \frac{\text{number of permutations where } |D_{\text{perm}}| \geq |D_{\text{obs}}|}{N},$$

where D_{obs} is the observed difference in GRADE scores between the two models, and D_{perm} is the difference obtained under each permutation. We compare the p-value p to a significance level of $\alpha = 0.05$.

Results. The vast majority of pairs are statistically significant.

Comparisons based on single-prompt distributions reveal just three pairs are not statistically significant: (SDXL, SDXL-LCM), (SDXL, DeepFloyd-XL), and (SDXL-Turbo, FLUX.1-schnell).

Similarly, comparisons using multi-prompt distributions, reveal only 15 pairs are not statistically significant:
(SD-1.1, SD-1.4), (SD-1.1, SD-2.1), (SD-1.1, DeepFloyd-M), (SD-1.1, DeepFloyd-L), (SD-1.4, SD-2.1), (SD-1.4, DeepFloyd-M), (SD-1.4, DeepFloyd-L), (SD-2.1, DeepFloyd-M), (SD-2.1, DeepFloyd-L), (SD-2.1, DeepFloyd-XL), (SDXL, SDXL-LCM), (DeepFloyd-L, DeepFloyd-XL), (SD-3 (2B), FLUX.1-schnell), (SD-3 (2B), FLUX.1-dev), and (FLUX.1-schnell, FLUX.1-dev).

Non-significant pairs are similar in quality. For example, all pair combinations of SD-1.1, SD-1.4, and SD-2.1 are not significant, which is not surprising since these models largely share the same underlying architectures and training data.

Why standard deviation can be misleading. We report standard deviations in Table 13 for completeness, but emphasize that they can be uninformative for multi-modal or heavily skewed distributions. A single standard deviation hides whether most values cluster around a single region or split between two (or more) distinct clusters, producing a deceptively large overall variance. Indeed, in Figure 15, many models show a pronounced high-low split in their GRADE scores.

	GRADE Score ↑	
Model	**Multi-prompt**	**Single-prompt**
DeepFloyd-M	$\mathbf{0.64 \pm 0.30}$	0.49 ± 0.34
DeepFloyd-L	0.62 ± 0.29	0.47 ± 0.34
DeepFloyd-XL	0.61 ± 0.30	0.46 ± 0.34
SD-1.1	$\mathbf{0.64 \pm 0.30}$	$\mathbf{0.54 \pm 0.33}$
SD-1.4	$\mathbf{0.64 \pm 0.29}$	0.53 ± 0.33
SD-2.1	0.63 ± 0.30	0.51 ± 0.34
SDXL	0.59 ± 0.31	0.46 ± 0.34
SDXL-Turbo	0.52 ± 0.33	0.36 ± 0.33
SDXL-LCM	0.58 ± 0.32	0.45 ± 0.34
SD-3 (2B)	0.47 ± 0.33	0.34 ± 0.33
FLUX.1-schnell	0.48 ± 0.33	0.36 ± 0.33
FLUX.1-dev	0.47 ± 0.33	0.32 ± 0.32

Table 13: **GRADE score in multi- and single-prompt distributions.** The mean entropy over all distributions for each model over multi-prompt and single-prompt settings. All models have a standard error of $\hat{\sigma} < 0.02$ and $\hat{\sigma} < 0.001$ respectively. Values close to 1 indicate highly diverse behavior (uniform) while values close to 0 indicate highly repetitive generations. The *most* diverse models are in bold.

D.3 Discussion of results

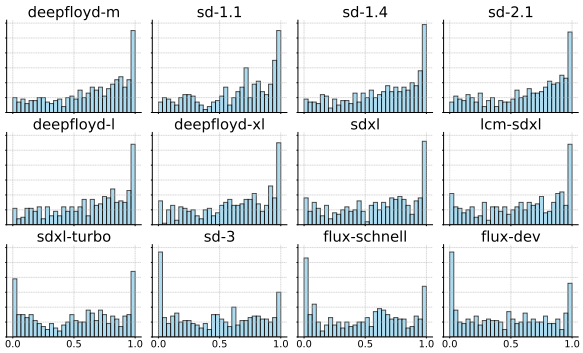
Our findings reinforce the observations made in the main text regarding the interplay between model scale, diversity, and prompt adherence:

Inverse-scaling law. There is a negative correlation between diversity and model size, suggesting that increasing model parameters leads to decreased diversity. This phenomenon is most apparent within each model family and aligns with the concept of an inverse-scaling law.

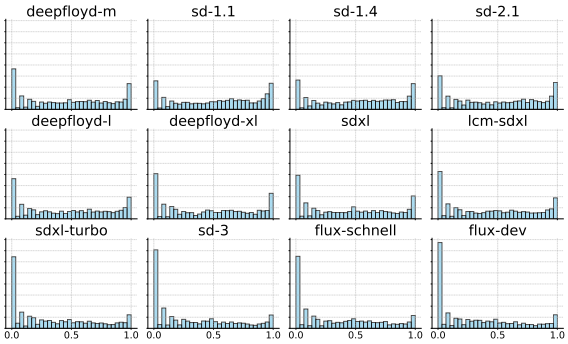
Fidelity-diversity trade-off. The negative correlation between diversity and prompt adherence indicates a trade-off between a model's ability to generate images that match the prompt and the diversity of its outputs. This is consistent with previous findings on fidelity-diversity trade-offs (Dhariwal and Nichol, 2021; Kynkäänniemi et al., 2019), where improving a model's prompt-adherence reduces the overall diversity of its outputs.

E Default Behaviors

In Section 4.1 we define default behaviors and mention that almost all concepts are associated with at least one default behavior, as shown in Section E. In Section E, we report the percentage of default behaviors



(a) **Histogram from each model in the multi-prompt setting.**



(b) **Histogram from each model in the single-prompt setting.**

Figure 15: **A histogram of the GRADE scores (normalized entropy) from each model for both distribution types.** Except the histograms of the most diverse models in the multi-prompt setting, histograms exhibit bimodal distributions, with peaks near both tails.

for both types of distributions.

Section E shows a sample of default behaviors detected in multi-prompt distributions and Figure 16 images of these behaviors.

	% of Default Behavior ↓	
Model	**Multi-prompt**	**Single-prompt**
DeepFloyd-M	83	92
DeepFloyd-L	81	92
DeepFloyd-XL	80	92
SD-1.1	78	87
SD-1.4	82	87
SD-2.1	76	89
SDXL	81	90
SDXL-Turbo	86	95
SDXL-LCM	82	92
SD-3 (2B)	88	95
FLUX.1-schnell	**90**	**97**
FLUX.1-dev	88	96

Table 14: **Percentage of at least one default behavior.** Lower values indicate higher diversity. Almost all concepts are associated with at least one default behavior in single prompt distributions, with a similar trend in multi-prompt distributions. The model with the *most* default behaviors is in bold.

Model	% of Default Behavior ↓	
	Multi-prompt	Single-prompt
DeepFloyd-M	39	54
DeepFloyd-L	39	56
DeepFloyd-XL	40	56
SD-1.1	39	49
SD-1.4	40	51
SD-2.1	40	52
SDXL	44	57
SDXL-Turbo	50	67
SDXL-LCM	44	57
SD-3 (2B)	56	69
FLUX.1-schnell	55	67
FLUX.1-dev	**56**	**70**

Table 15: **Percentage of all default behaviors.** Lower values indicate higher diversity. There are 405 multi-prompt and 2430 single prompt distributions in total. The table quantifies the total percentage of default behaviors observed. The model with the *most* default behaviors is in bold.

Model	Question (Attribute)	Attribute Value	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

Table 16: **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 Figure 16.

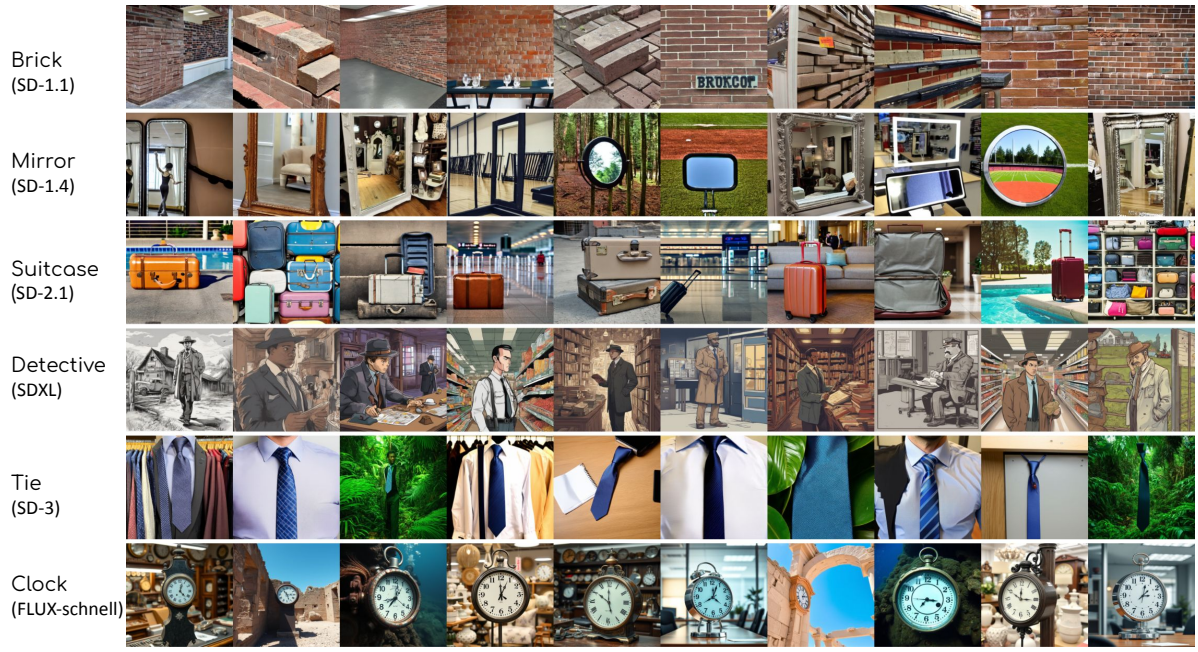


Figure 16: **A sample of images depicting the default behaviors in Section E.** The concept is shown in the left column with the model directly below it. Images were sampled randomly from all prompts. The default behaviors, top down: (1) stacked bricks; (2) framed mirrors; (3) hard-shell suitcase; (3) male detective; (4) neckties; and (5) analog clocks.

F Low Diversity Originates in Training Data

Filtering Captions from LAION. We aimed to measure the diversity of training images whose captions satisfy two conditions: (1) they mention the concept as an object and not as a modifier (e.g., "cookie" but not "cookie cutter"), and (2) the caption must not mention or imply the attribute of interest (e.g., "a classic chocolate chip cookie" implies the cookie is round). We queried LAION using WIMBD (Elazar et al., 2024) and sampled 500 captions for each concept.

To efficiently filter the captions, we utilized GPT-4o in a few-shot setup. For each caption, we provided the caption text, the concept (e.g., "cookie"), and the question regarding the attribute of interest (e.g., "what is the shape of the cookie?"). We instructed GPT-4o to analyze each caption and determine whether it satisfies both filtering conditions. The model was prompted to reply with "yes" if both conditions are met and "no" otherwise.

We then downloaded the images associated with the captions that GPT-4o classified as satisfying both conditions. To ensure the reliability of our filtering method, we conducted a human evaluation, achieving an F1 score of 90.3%. Detailed methodology and results of the human evaluation are provided in Section H.

Below is the prompt we use with GPT-4o to filter captions from LAION:

```
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. The question asks about the color of the chair, and there is no mention of a chair 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
```

G Validating GRADE

We validate GRADE with five targeted experiments. We separate *data construction checks* (E1–E3) from *human/crowd validation* (E4–E5). Step (d) of GRADE (normalized entropy) is purely formulaic and requires no empirical validation. Across studies in this section, we use 2,800 images sampled from 12 T2I models. These one-time checks establish the soundness of GRADE; they are not required prior to every use.

Data construction checks

E1 — Prompt concept only (no modifiers). We manually review all 600 prompts and confirm each explicitly mentions the intended concept and does not contain attribute descriptors. As such, no prompt contains attribute descriptors.

E2 — Common vs. uncommon prompts. To verify our split, we extract all nouns from each prompt and measure their co-occurrence in LAION-5B (Schuhmann et al., 2022) using WIMBD (Elazar et al., 2024). We find that nouns in *common* prompts co-occur on average 30,655 times vs. 956 in *uncommon* prompts, confirming a large distributional gap between the two. E.g., there are 49,697 prompts with both "bear" and "forest" in LAION-5B, and only 141 prompts with "bear" and "shopping mall".

E3 — Attribute validity. We manually verify that each of the 405 questions generated by GRADE is about a visual aspect of the target concept and can be answered by viewing images showing the concept. For example, the *hairstyle* of a princess is visually verifiable, but not her personality. We also verify that no two attribute values in the answer set semantically overlap (e.g., "round" vs. "circle"). Indeed, we find that all questions are visually grounded and answer sets do not overlap.

Human / crowd validation

To validate that GRADE produces reliable VQA judgments and that its attribute value sets are complete, we conduct a large-scale AMT crowdsourcing study with 71 experienced workers (over 5,000 approved HITs and over 98% approval rate) who passed a dedicated qualification exam, taking the majority decision from three independent judgments for each case.

E4 — VQA–human agreement (AMT). We evaluate GPT-4o as the VQA backbone in two studies. (i) Broad: 1,000 images across 12 models—agreement with human majority selection is **90.2%**. (ii) Focused: a multi-prompt distribution with 1,800 images from SD-1.1, FLUX.1-dev, and SDXL—overall agreement is **92.8%**, per-model: 88.0%, 91.2%, and 99.5%, respectively. More detail in Section H.

E5 — "None of the above" handling. The VQA in GRADE is instructed to mark unanswerable image-question pairs with "none of the above", either because the concept is absent from the image–a

failure of the image generation model–or the answer falls outside the set of generated attribute values, making it insufficient. Here, we verify that while such cases are infrequent, GRADE handles them well. Of the 1,000 images in the first study in E4, only 115 examples were marked as unanswerable either by human selection or the VQA. 92 are images that omit the concept (mostly images by SD-1.1, which is notoriously weak in prompt adherence). In only three cases the correct answer was missing from the answer set. Finally, the other 20 are disagreements between human selection and the VQA, most of which are indeed equivocal. For example, the bottom right image for "Is the person male or female?" in Figure 20. This confirms that GRADE reliably flags genuine out of scope cases.

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% (as well as be native-level English readers and writers). 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. In total, 71 unique workers participated in evaluating GRADE and 49 to filter the captions from LAION.

Validating GRADE. To validate the VQA in Section 5, we run an AMT crowdsourcing task where each worker is provided with a question, concept, image, and attribute values, and is requested to select the attribute value that best matches the question and image. The UI for this task can be viewed in Figure 17 with examples in Figure 18. A sample of cases from our attribute values coverage validation (validation of step (b)) is available in Figure 19 and Figure 20.

Validating filtering of captions from LAION. To assess the effectiveness of our GPT-4o-based caption filtering method described in Section 6, we conducted an Amazon Mechanical Turk (AMT) crowdsourcing task. We sampled 1,000 captions from LAION, ensuring an equal distribution of 500 captions that met the filtering criteria and 500 that did not. Workers were instructed to evaluate whether each caption (1) explicitly mentioned the concept as the main object rather than as a modifier (e.g., "cookie" instead of "cookie cutter") and (2) the caption must not mention or imply the attribute of interest (e.g., "a classic chocolate chip cookie" implies the cookie is round). Each example was reviewed by three independent workers, and the majority decision was taken as the final label. Our automated filtering method achieved a recall of 85.8% and a precision of 95.4%, resulting in an F1 score of 90.3%, which indicates a high level of agreement with human judgments. These findings demonstrate that GPT-4o is a reliable tool for automated caption filtering. Additional details about the user interface and example cases are provided in Figure 21 and Figure 22, respectively.

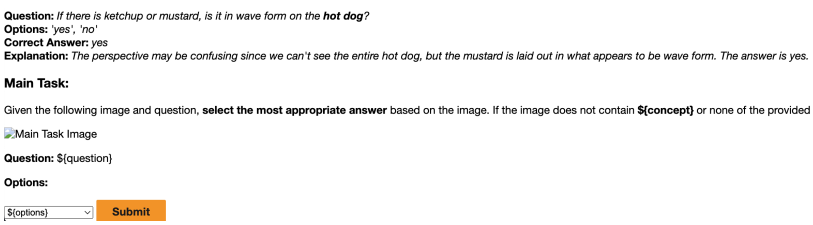


Figure 17: A screenshot of the VQA validation task. Workers are provided a question, concept, image, and a set of categories, including "none of the above" (options here). Their task is to select the option that answers the question.

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 18: 3 out of 10 examples provided to workers as aid to complete their visual question answering task.

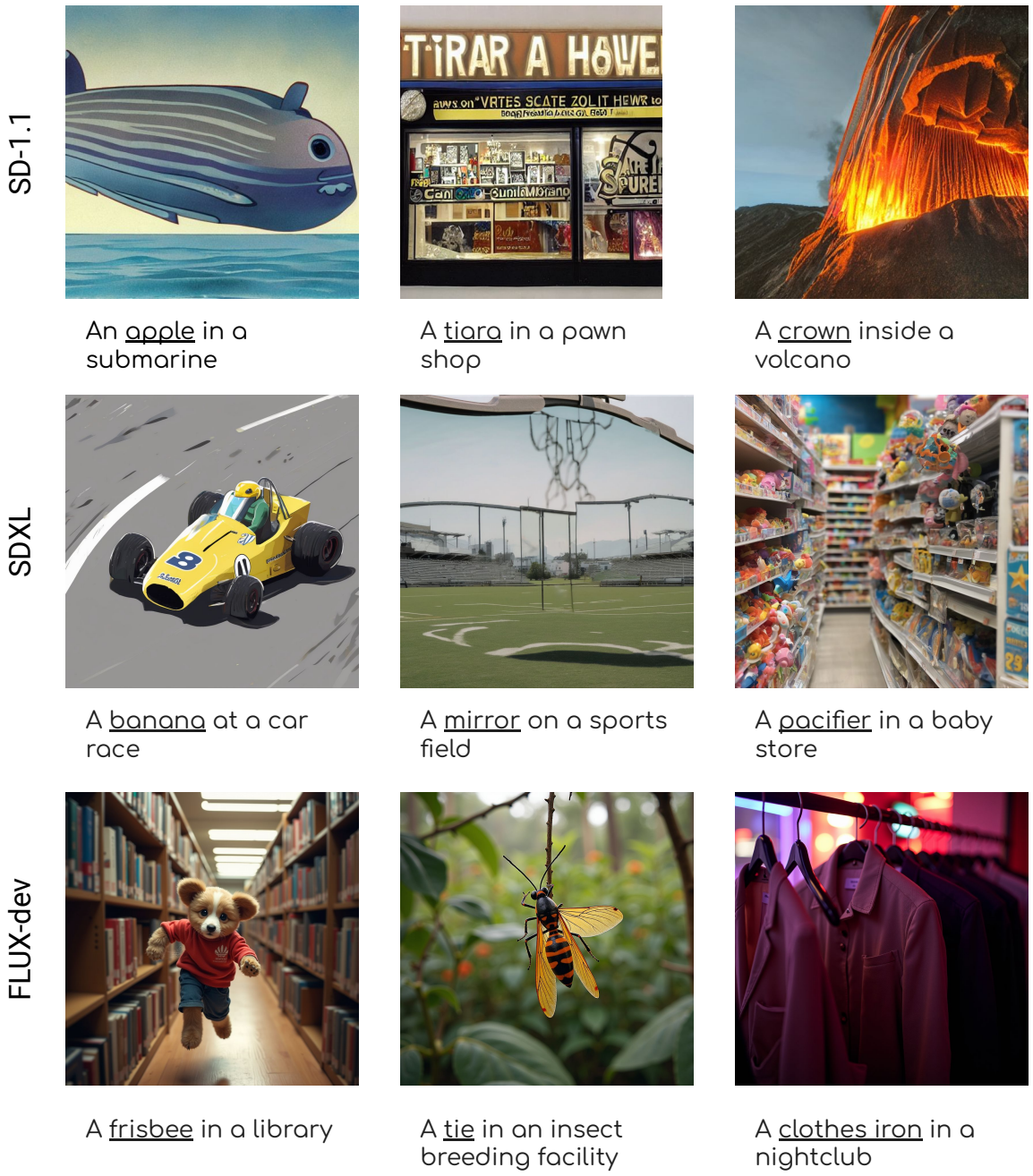


Figure 19: A sample of images marked with “none of the above”, as a result of not including the concept (underlined) in the image.

SDXL

Popcorn at a cinema

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

V^a_c = {bucket, bowl}

SD-1.1

a toy at a children's playroom

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

V^a_c = {mechanical, electronic}

SDXL

a tie in an office

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

V^a_c = {casual, formal}

FLUX-dev

A person in a city square

Q: Is the person male or female?

V^a_c = {male, female}

Figure 20: A sample of images marked with "none of the above". The top row exhibits cases where the attribute value is not in $\mathcal{V}_c^a$. The bottom row exhibits cases where the question cannot be answered just from viewing the image. The concept in each prompt is underlined.

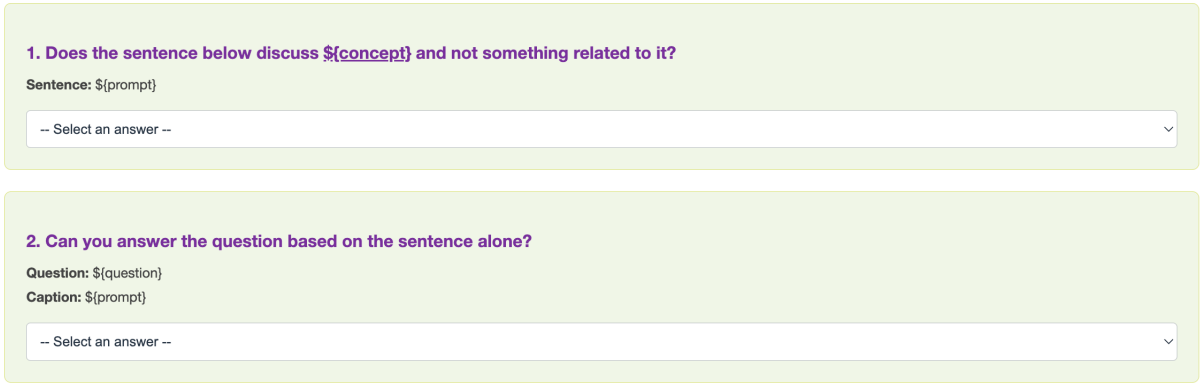


Figure 21: A screenshot of the caption filtering validation task. Workers are provided a caption, two questions, and a concept. Their task is to read the caption and answer the questions.

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 22: 3 out of 10 examples provided to workers as aid to complete their caption filtering task.

I Prompts in GRADE

I.1 Concept Collection

To collect a list of diverse concepts, we prompt GPT-4o (OpenAI et al., 2024) 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
```

I.2 Prompt generation

The following prompt was used to generate common 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}
```

This one was used to generate uncommon 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}
```

I.3 Attribute Generation

GRADE first analyzes the specific attributes of the concept provided in the prompt, and then generates questions that can be used to count the occurrences of attribute values in images. Below is the prompt we used with GPT-4o.

```
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}
```

I.4 Attribute Values Generation

To generate attribute values $\tilde{\mathcal{V}}_c^a$ for $\tilde{P}_{V|a,c}$, we provide GPT-4o (OpenAI et al., 2024) with a concept, a question, and a prompt. GPT-4o then outputs a list of attribute values that can match the question (attribute). The process is performed for all prompts mentioning 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 $\tilde{\mathcal{V}}_c^a$.

```
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 first analyze the description of the image and the question, then, hypothesize
↪  plausible answers that can surface from viewing the image. Do not write anything other than
↪  the answer.
Then, I need you to list the plausible answers in a list, just like in the example below. For
↪  example,
Caption: a helmet in a bike shop
Question: What type of helmet is depicted in the image?
Plausible answers: ["motorcycle helmets",
            "bicycle helmets",
            "football helmets",
            "construction helmets",
            "military helmets",
            "firefighter helmets",
            "rock climbing helmets",
            "hockey helmets"]
Now your turn.
Caption: {caption}
Question: {question}
Plausible answers:


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 first carefully read the question and analyze, then hypothesize
↪  plausible answers to the question assuming you could examine the image (instead, you examine
↪  the caption). The answers should be in a list, as in the example below. Do not write anything
↪  other than the plausible answers.

Example:
Caption: a helmet in a bike shop
Question: What type of helmet is depicted in the image?
Plausible answers: ["motorcycle helmets",
            "bicycle helmets",
            "football helmets",
            "construction helmets",
            "military helmets",
            "firefighter helmets",
            "rock climbing helmets",
            "hockey helmets"]
Now your turn.
Caption: {caption}
Question: {question}
```

```
Plausible answers:
```

I.5 Generating answers

We use GPT-4o to answer the generated questions with $1,000$ as max tokens and temperature 0. We use the Structured Outputs feature (OpenAI, 2024) to map the natural language answers to attribute values in a single step. 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:
```