Quantifying and Mitigating Default Behaviors in Text-to-Image Models

Anonymous ACL submission

Abstract

Text-to-image (T2I) models consistently collapse to stereotypical attribute values when prompts leave concepts underspecified, a phenomenon we term *default behaviors*. We conduct a large-scale empirical study using GRADE, an LLM-driven measurement approach that quantifies diversity at the level of individual concept attributes. We find pervasive default behaviors across all evaluated T2I models, with stronger models tending to be less diverse. We trace these defaults to reporting bias in training captions: when captions omit common attributes, the accompanying images are dominated by a narrow set of values, and model outputs closely mirror this skew. We further show that a simple mitigation technique, making underspecified attributes explicit in the prompt, is sufficient to mitigate this collapse.

1 Introduction

Text-to-image (T2I) models generate remarkably realistic images from textual descriptions, yet prompts are inherently *underspecified* (Hutchinson et al., 2022; Rassin et al., 2022); they rarely constrain a concept to a specific subtype. The phrase "a princess at a children's birthday party" does not specify the princess's outfit, skin color, or age. In principle, this openness should let the model sample across the wide range of valid realizations. In practice, however, models consistently depict her as a white girl in a pink dress and tiara (Figure 1). We call this systematic collapse *default behavior*: the tendency of a model to repeatedly generate the same narrow set of attribute values for an underspecified concept. How pervasive are default behaviors, and how can we mitigate them?

Feature-level metrics such as FID (Heusel et al., 2017), Precision-and-Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), and more recently, Vendi Score (Friedman and Dieng, 2023) measure diversity in embedding space. Because they operate in representation space, they cannot isolate *which* attributes lack variation, making it difficult to diagnose the source of low diversity.

To study default behaviors at the attribute level, we use **GRADE** (**Gr**anular **A**ttribute **D**iversity **E**valuation), a measurement approach that leverages an LLM to derive concept-specific attributes and a VQA model to extract attribute values from generated images, yielding categorical distributions whose normalized entropy quantifies diversity (illustrated in Figure 2). Using GRADE, we conduct a large-scale empirical study across 12 T2I models and over 720K images. Our findings reveal:

- **Pervasive default behaviors:** All 12 models exhibit low diversity, with stronger models consistently scoring lower. The most diverse model scores only 0.64, and default behaviors occur in up to 97% of concepts.
- **Reporting bias in training captions:** We find strong correlational evidence that default behaviors originate in training data. In LAION, when captions omit an attribute, the accompanying images are dominated by a narrow set of values, and Stable Diffusion outputs closely mirror this skew (Pearson $r \approx 0.8$). This connects default behaviors to *reporting bias* (Gordon and Van Durme, 2013), where common attributes go unmentioned.
- **Mitigation by prompt rewriting:** We demonstrate a simple LLM-based, attribute-focused rewriter that samples concept attributes and injects them into underspecified prompts, nearly doubling FLUX.1-dev's GRADE score without any modification to the T2I model.

2 Related Work

Diversity and bias in T2I models. A growing body of work documents systematic biases and lack of diversity in T2I outputs. Large-scale audits have

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

shown that models default to stereotypical depictions across gender (Girrbach et al., 2025), race (Luccioni et al., 2023), and even object attributes (Choi et al., 2025). Seshadri et al. (2023) show that apparent bias amplification partly stems from distribution shifts between training captions and evaluation prompts, rather than the model blindly amplifying training statistics. Simonen et al. (2026) study "default images" in Midjourney, finding that vague prompts trigger visually similar outputs. Our work differs in two ways: we measure defaults at the *attribute* level (e.g., the color of a bottle, not just the overall image), and we trace defaults to reporting bias in training captions rather than characterizing the phenomenon alone. Adamkiewicz et al. (2026) find that newer T2I models produce less diverse outputs despite improved fidelity.

Training data and caption quality. Several studies show that properties of training captions directly shape T2I model behavior. Brack et al. (2025) find that dense synthetic captions improve alignment but can hurt diversity, and that caption distributions shift downstream output bias. Cheng et al. (2024) study which caption properties matter for T2I training, showing that incomplete or inaccurate captions degrade generation quality. Seshadri et al. (2024) show that the ratio of visual depiction to textual mention of entities in training data correlates with generation success. At the VLM level, Al Sahili et al. (2025) find that data source is a stronger driver of bias than model size.

Diversity metrics. Classical diversity metrics such as FID (Heusel et al., 2017) and Precision-and-Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) compare feature-level distributions against a reference set. However, they cannot isolate which attributes lack variation and are sensitive to data contamination (Kynkäänniemi et al., 2022). Vendi Score (Friedman and Dieng, 2023; Pasarkar and Dieng, 2023) is reference-free and computes a single holistic diversity number from a similarity matrix in embedding space. While general and theoretically grounded, it cannot decompose diversity into interpretable attributes or identify *which* visual property collapsed. Our empirical comparison (Appendix C) shows that GRADE

agrees with embedding-space metrics on coarse model rankings but captures attribute-level variation that they miss. OpenBias (D'Incà et al., 2024) is closest to our pipeline: it also uses LLMs to propose evaluation dimensions and VQAs to assess generated images, quantifying distributional skew via normalized entropy. However, the two methods differ in research question, experimental design, and scope. OpenBias detects social biases to assess *fairness*, whereas GRADE measures diversity collapse across general visual attributes for underspecified concepts. OpenBias evaluates three Stable Diffusion variants using existing COCO/Flickr30k captions, while GRADE evaluates 12 models across 100 concepts using deliberately underspecified prompts that isolate default behaviors. Most distinctively, GRADE traces diversity collapse to reporting bias in training captions and mitigates it, going beyond bias detection.

3 Measurement Methodology

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.

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 measurement approach comprises four steps, illustrated in Figure 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 6 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?").

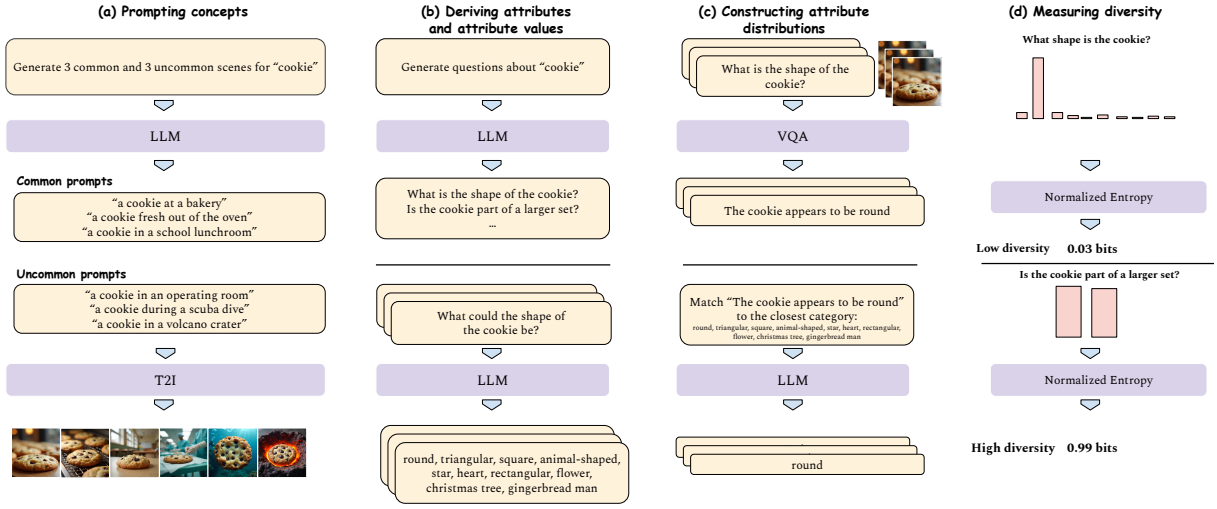


Figure 2: **Workflow of GRADE using "cookie" as input.** (a) Generate prompts that mention "cookie" without specifying its attributes, and use them to generate images. (b) Formulate attribute-related questions and extract responses from the images using a VQA model. (c) Produce attribute values and map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

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 5 provides a representative sample of the generated concepts, attributes, and attribute values, illustrating how $\tilde{\mathcal{V}}_c^a$ is constructed.

(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.

We use GPT-4o (OpenAI et al., 2024) as the LLM and VQA backbone. For each concept, we generate three common and three uncommon prompts, each yielding 100 images, with four attributes per concept on average. Full implementation details, prompt templates, and cost analysis are provided in Appendix I. Sample concepts, attributes, and prompts are shown in Appendix B.

4 Default Behaviors in Text-to-Image Models

We apply our measurement approach to 12 T2I models and find pervasive default behaviors across all models.

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, yielding 405 multi-prompt and 2,430 single-prompt distributions (60,000 images/model).

T2I models. We evaluate 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-

Model	GRADE Score ↑	
	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 1: **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.

Model	% of Default Behavior ↓	
	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 2: **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.

dev. All models use default `Diffusers` (von Platen et al., 2023) settings.

4.1 Results

All models have low diversity scores. Table 1 presents mean GRADE scores of models across multi- and single-prompt distributions. Permutation tests confirm the observed differences are statistically significant (see Appendix 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 Appendix A.

Stronger models tend to be less diverse. The lowest GRADE scores in Table 1 are produced by FLUX.1-dev, FLUX.1-schnell, and SD-3 (2B).

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. Appendix E includes aggregate results for models.

Results. We find that even the distributions from the **most** diverse model (SD-1.1) collapse **87%** of the time in a single-prompt setting, and **76%** in a multi-prompt setting. Complete results with further analyses are provided in Appendix 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 Measurement Reliability

Since our findings depend on the measurement pipeline, we validate both data quality and VQA reliability (full details in Appendix G). For data quality, we verify that all 600 prompts mention the target concept without specifying attributes, and confirm the common/uncommon split via noun co-occurrence in LAION-5B (30,655 vs. 956 on average). All attribute questions are visually grounded with non-overlapping answer sets.

For VQA reliability, we use 71 qualified AMT workers to answer the same questions over 2,800 images (3-way majority vote). The GPT-4o-based VQA agrees with human judgments 91.8% of the time. Of 115 flagged images, 92 were cases where the T2I model failed to depict the concept (a model failure, not a measurement failure), only 3 had incomplete attribute value sets (measurement failure), and 20 were genuinely ambiguous cases.

We compare GRADE to FID, Recall, and Vendi Score in Appendix C: FID and Recall correlate near-zero with GRADE, and Vendi Score correlates only weakly per concept (r=0.26–0.34), so neither recovers attribute-level diversity.

6 Reporting Bias in Training Captions

The previous section established that T2I models exhibit pervasive default behaviors. We now inves-

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.

tigate their origin. Our hypothesis is that default behaviors stem from *reporting bias* (Gordon and Van Durme, 2013) in training captions: when a concept is mentioned without an explicit attribute value (e.g., "umbrella" rather than "open umbrella"), the accompanying images in the dataset are dominated by a narrow set of attribute values. The model then learns to replicate this skewed distribution whenever it encounters an underspecified prompt.

This hypothesis is grounded in a well-known linguistic phenomenon: speakers systematically omit attributes they consider "obvious" or "typical" (Gordon and Van Durme, 2013). In the context of web-scale image-caption datasets like LAION, this means captions rarely specify whether an umbrella is open or closed, or the shape of a cookie. These are precisely the attributes that collapse to defaults in model outputs. To test this, we compare attribute distributions of real LAION images whose captions omit an attribute against model generations from similar captions.

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 Appendix 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 3 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 3), 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.

These results are consistent with our hypothesis: when concept-attribute pairs are left unspecified in captions, the accompanying images are dominated by a narrow set of attribute values, and model outputs closely mirror this skew. The strong correlations suggest that reporting bias in training captions is a key factor associated with default behaviors. Unless the user explicitly specifies an attribute, the model reproduces the distribution it has observed most frequently in training, resulting in a systematic lack of diversity.

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

7 Mitigating Default Behaviors via Prompt Rewriting

Default behaviors arise when prompts leave attributes unspecified. A natural mitigation is therefore to remove the underspecification: sample plausible attribute values and write them into the prompt before generation. Closest to our approach, Datta et al. (2024) train a model that expands user queries into more diverse text prompts for T2I generation. We pair GRADE's concept-attribute taxonomy with a simple prompt rewriter: given an underspecified prompt, it samples attribute values and injects them to the prompt.

7.1 Method

The rewriter draws its sampling distribution from a *concept store*, populated by an LLM: the first time a concept is seen, an attribute list with candidate values is generated and cached; subsequent generations reuse it. For an input prompt p and a budget of N generations, the rewriter repeats the loop:

1. **Concept extraction.** An LLM extracts the concepts in p and their store, creating them on demand as described above.

2. **Attribute sampling.** The rewriter samples $k \in \{0, \ldots, K\}$ `(attribute, value)` tuples uniformly from the loaded entries (e.g., `species=panda`, `pose=sitting`); when $k = 0$, the prompt passes through unchanged.

3. **Minimal rewriting.** A second LLM call rewrites p to include the sampled tuples with the minimum lexical change required for grammaticality.

Concept	GRADE Score ↑ Base	Rewriter	Δ
a bear	0.33	0.74	+0.42
cheese	0.46	0.65	+0.19
a cow	0.52	0.91	+0.40
a pacifier	0.35	0.63	+0.29
a ring	0.48	0.74	+0.26
a spinning wheel	0.42	0.70	+0.28
a vase	0.47	0.81	+0.33
coffee	0.26	0.77	**+0.51**
soap	0.26	0.54	+0.28
yogurt	0.31	0.58	+0.27

Table 4: **Base vs. rewriter.** Higher is better. Mean across 228 single-prompt distributions: 0.36 vs. 0.69.

4. **Generation.** The rewritten prompt is sent to the frozen T2I backend.

For example, given p = *"a bear in a forest"* and the sampled tuples `species=panda` and `pose=sitting`, the rewriter produces *"a sitting panda in a forest"*, which the T2I backend then renders. Implementation details are in Appendix J.

7.2 Experimental Setup

We pair our rewriter with FLUX.1-dev on 10 GRADE concepts (all 6 prompts each; $N = 100$ rewritten-prompt images per prompt, matching Section 3), yielding 6,000 images, and compare against FLUX.1-dev's baseline single-prompt scores. The concept store is seeded with GRADE's taxonomy so the rewriter samples along the axes GRADE evaluates.

7.3 Results

Our rewriter raises FLUX.1-dev's mean normalized entropy from 0.36 to 0.69 across 228 single-prompt distributions.

The lift concentrates on attributes that collapse under the baseline.Bear *species* for the prompt *"a bear in a forest"* is a clean case: FLUX.1-dev produces a grizzly bear in nearly every generation, while the rewriter still produces grizzlies in most images but covers all six bear species in the value set. Qualitative grids in Figure 4 demonstrate that rewriter samples remain on-prompt while spanning visibly more of the attribute space. This demonstrates that default behaviors admit a simple inference-time mitigation: a small LLM loop around a frozen T2I model substantially shifts its output distribution.

8 Conclusion

Our large-scale analysis of 12 T2I models reveals that default behaviors are pervasive across all models we tested. This represents a blind spot in current evaluation practices: because sample diversity is not systematically measured, default behaviors persist unchecked across model generations. We trace these defaults to reporting bias in training captions. When captions omit common attributes, the accompanying images are skewed, and models faithfully reproduce this skew.

Figure 4: Samples for two underspecified prompts. FLUX.1-dev (top row of each block) collapses to a single mode: a brown grizzly and white yogurt in a white bowl. Our rewriter (bottom row) spans clearly distinct attribute values (bear species; yogurt color) while remaining on-prompt.

This connection between a well-known linguistic phenomenon (what is said and unsaid in text) and a concrete failure mode of visual generation suggests that improving diversity requires attending to the training data itself, particularly the way concepts are described in captions. We further demonstrate that simply making underspecified attributes explicit in the text is sufficient to mitigate this collapse.

These results have implications for both model evaluation and training practices. We encourage future work on diversity-aware training objectives, caption enrichment strategies that reduce reporting bias, and systematic diversity benchmarking.

Limitations

Our analysis has two main limitations. First, the findings are scoped to the concepts and attributes we measure; attributes not included in the evaluation remain unassessed, and the picture of diversity may shift with a different selection. Second, our measurement pipeline relies on an LLM and a VQA model, which may introduce their own biases into the attribute-suggestion and answer-extraction steps. While our human validation (91.8% agreement) suggests these biases are limited, they cannot be fully ruled out.

Ethics Statement

Our work measures default behaviors and diversity in text-to-image models. We do not introduce new models or datasets. Our measurements reveal existing patterns rather than creating new ones, and we hope they contribute to more equitable model development.

References

Krzysztof Adamkiewicz and 1 others. 2026. When pretty isn't useful: Investigating why modern text-to-image models fail as reliable training data generators. *arXiv preprint arXiv:2602.19946*.

Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. 2025. Data matters most: Auditing social bias in contrastive vision-language models. *arXiv preprint arXiv:2501.13223*.

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.

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.

Manuel Brack, Sudeep Katakol, Felix Friedrich, Patrick Schramowski, Hareesh Ravi, Kristian Kersting, and Ajinkya Kale. 2025. How to train your text-to-image model: Evaluating design choices for synthetic training captions. *arXiv preprint arXiv:2506.16679*.

Sheng Cheng, Maitreya Patel, and Yezhou Yang. 2024. Precision or recall? an analysis of image captions for training text-to-image generation model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3703–3709. Association for Computational Linguistics.

Dasol Choi and 1 others. 2025. When cars have stereotypes: Auditing demographic bias in objects from text-to-image models. *arXiv preprint arXiv:2508.03483*.

Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. 2024. Prompt expansion for adaptive text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

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. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

Leander Girrbach, Stephan Alaniz, Giles Smith, and Zeynep Akata. 2025. A large scale analysis of gender biases in text-to-image generative models. *arXiv preprint arXiv:2503.23398*.

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.

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*.

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.

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.

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, Yasaman Razeghi, Sameer Singh, and Yanai Elazar. 2024. The association between training data and text-to-image generation capabilities. In *NeurIPS 2024 Workshop on Attributing Model Behavior at Scale (ATTRIB)*.

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

Hannu Simonen, Atte Kiviniemi, Hannah Johnston, Helena Barranha, and Jonas Oppenlaender. 2026. An exploration of default images in text-to-image generation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM.

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

This appendix provides qualitative examples that complement the quantitative results in the main paper. Figures 5 and 6 contrast generations from SD-1.4, SDXL, and FLUX.1-dev on representative underspecified prompts, illustrating how diversity collapses as models grow stronger. Figures 7 to 9 then walk through the per-prompt GRADE pipeline on individual concepts, showing how the VQA-derived attribute distribution and its normalized entropy combine to produce a single diversity score.

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.

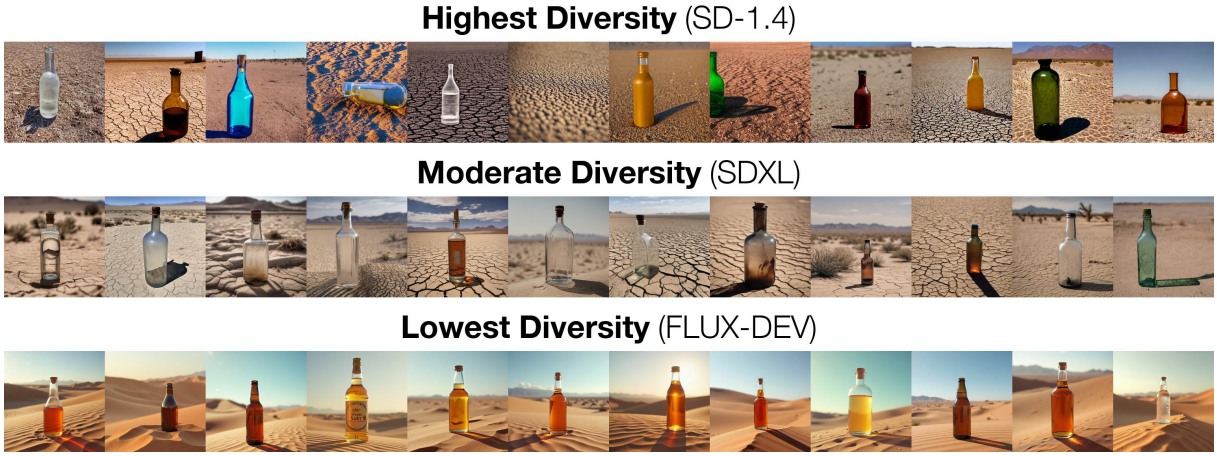


Figure 6: **Difference in diversity between models.** Images generated using the prompt "a bottle in a desert". Each row corresponds to a model, top-down: SD-1.4 (most diverse), SDXL, and FLUX.1-dev (least diverse). While no model exhibits high diversity, there is a marked difference between SD-1.4 and 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.

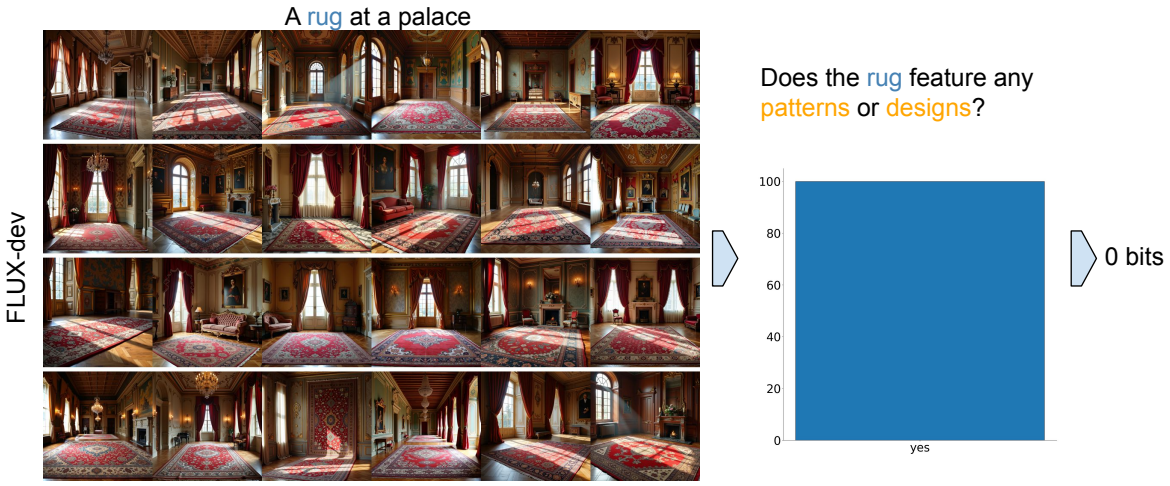


Figure 8: **Illustration of GRADE score.** Displayed are 24 of the 100 images generated by FLUX.1-dev using the prompt "A rug at a palace". The accompanying histogram and the subsequent entropy plot both represent the 100 sample. The GRADE score is 0, indicating the rugs are consistently patterned.

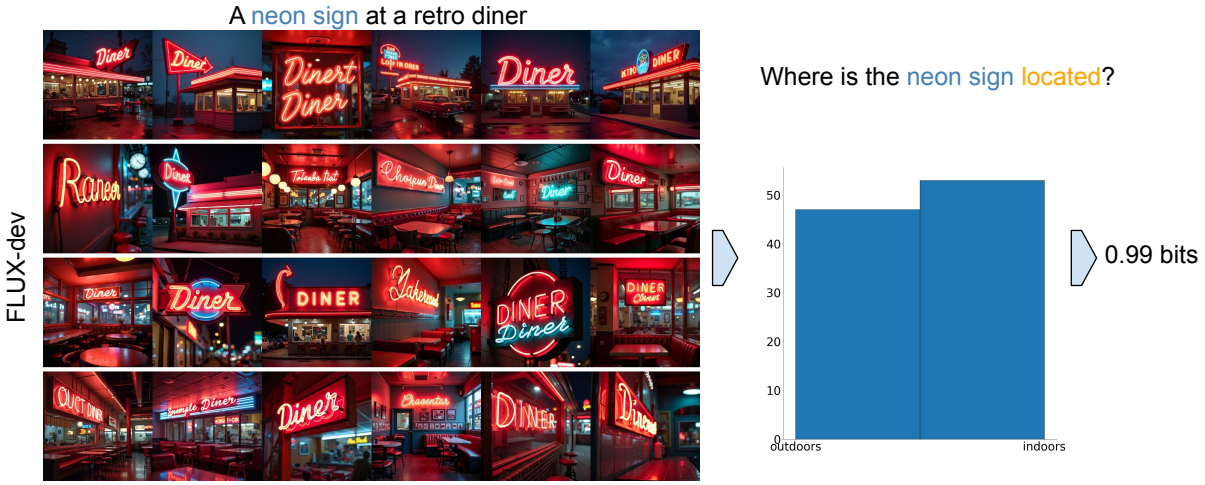


Figure 9: **Illustration of GRADE score.** Displayed are 24 of the 100 images generated by FLUX.1-dev using the prompt "A neon sign at a retro diner". The accompanying histogram and the subsequent entropy plot both represent the 100 sample. The GRADE score is 0.99, indicating the location of the signs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

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

B Extended Data Overview

This appendix details the data behind the GRADE benchmark. Table 5 shows a small sample of (concept, attribute, attribute-values) triplets to illustrate the structure of a multi-prompt distribution. Table 7 provides an expanded list of concept–attribute pairs covering shape, material, configuration, and other visual attributes, and Table 6 shows representative common and uncommon prompts used to elicit attribute-level diversity.

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 campsite.
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 6: **Sample concepts with common and uncommon prompts.**

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

C Comparing GRADE to Previous Metrics

Extended results for the comparison between GRADE and traditional metrics described in Section 5. Results using CLIP for feature extraction can be viewed in Table 8 and Table 9. Results using Inception v3 (Szegedy et al., 2014) (ImageNet features (Deng et al., 2009)) are in Table 10 and Table 11. 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 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

Figures 10 and 11 illustrate, on two representative concept–attribute pairs, why GRADE captures variation that FID and Recall miss. In both, the image sets are highly similar with respect to the target attribute (*wicker* picnic baskets; tablecloths without visible damage); TVD_G correctly reports this, while FID and Recall—which compare embedding distributions wholesale—read the same sets as dissimilar.

C.2 Comparison with Vendi Score

To further validate that attribute-level diversity differs from embedding-space diversity, we compare GRADE to Vendi Score (Friedman and Dieng, 2023), a reference-free metric defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix.

Setup. For each of the 100 concepts in the GRADE benchmark, we pool 50 generated images across the concept's six prompts and compute Vendi Score on CLIP ViT-B/32 cosine similarity, for a total of 5,000 images per model and 15,000 across all three models. The corresponding GRADE score is the mean attribute-level normalized entropy on the same concept. We evaluate three modern T2I models: SD-1.4, SDXL, and FLUX.1-dev.

VS and GRADE rank the models in the same order but correlate weakly per concept; individual concepts can diverge sharply. For *a tie* on SDXL, VS is high because the images vary in color and pattern, while GRADE collapses on two attributes: every tie is a necktie (never a bowtie) and worn with a formal outfit. Embedding-space variation hides an attribute-level default that attribute-centric methods like GRADE detect.

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.

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.

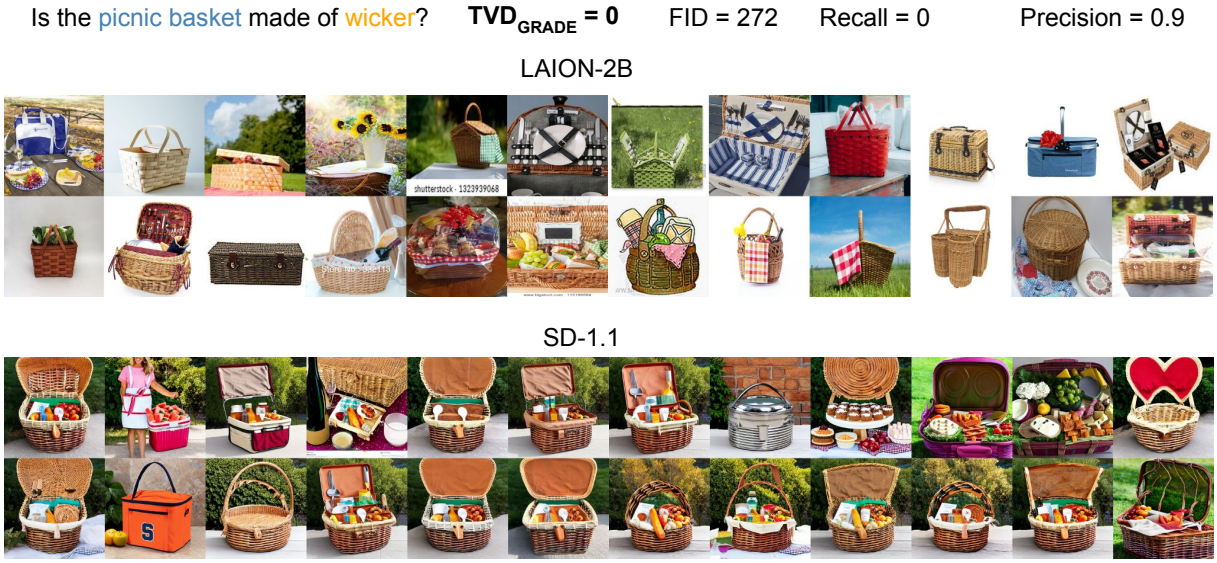


Figure 10: **Comparison between GRADE, FID, and Recall, using CLIP features.** The metrics are compared over the "wicker" attribute of the concept "picnic basket". $\text{TVD}_{\text{GRADE}}$ reports very high similarity (lower TVD is better) between the sets of images, which is indeed shown in the images (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.

Model	GRADE	VS	PCC	Spearman
SD-1.4	0.65	4.40	0.28	0.30
SDXL	0.60	4.04	0.26	0.21
FLUX.1-dev	0.48	3.43	0.34	0.35

Table 12: **GRADE and Vendi Score across 100 concepts per model.** VS uses CLIP ViT-B/32 cosine similarity over the multi-prompt image set; GRADE is the mean attribute-level normalized entropy. Both metrics rank the three models in the same order, but the per-concept correlation is only weak to moderate (PCC and Spearman, all $p < 0.025$).

D Extended Diversity Comparisons between T2I Models

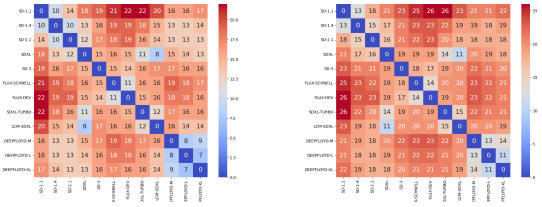


Figure 12: **The mean total variation distance (TVD) between all pairs of models over (a) multi-prompt distributions and (b) single prompt distributions.** For readability, both figures show TVD in a range between 0 and 100 instead of 0 to 1.

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

Table 13: **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 13.

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

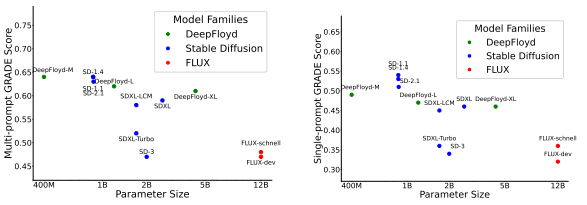


Figure 13: **(a) GRADE score in multi-prompt setting plotted against the denoiser's parameter size. (b) GRADE score in single-prompt setting plotted against the denoiser's parameter size.** To a degree, diversity deteriorates in tandem with parameter size. This phenomenon is most apparent within every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

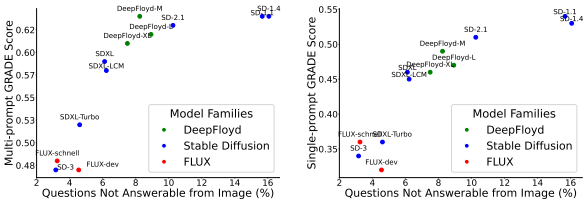


Figure 14: **(a) GRADE score in multi-prompt setting plotted against the % of "none of the above". (b) GRADE score in single-prompt setting plotted against the % of "none of the above".** In Section 5 we show 80% of which account for missing concepts in the image. The plots show negative correlation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

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 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 $\alpha = 0.05$.

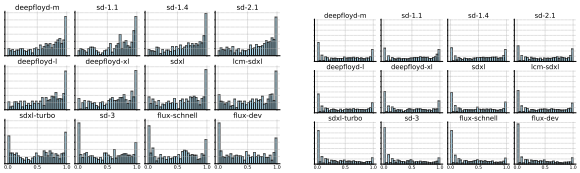
Results. The vast majority of pairs are statistically significant.

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 14 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.



(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.

	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 14: **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:

Larger models, less diverse. We observe a negative correlation between diversity and model size: increasing model parameters tends to decrease diversity. The pattern is most apparent within each model family.

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 Appendix E. In Appendix E, we report the percentage of default behaviors for both types of distributions.

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

Model	% of Default Behavior ↓	
	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 15: **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 16: **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.

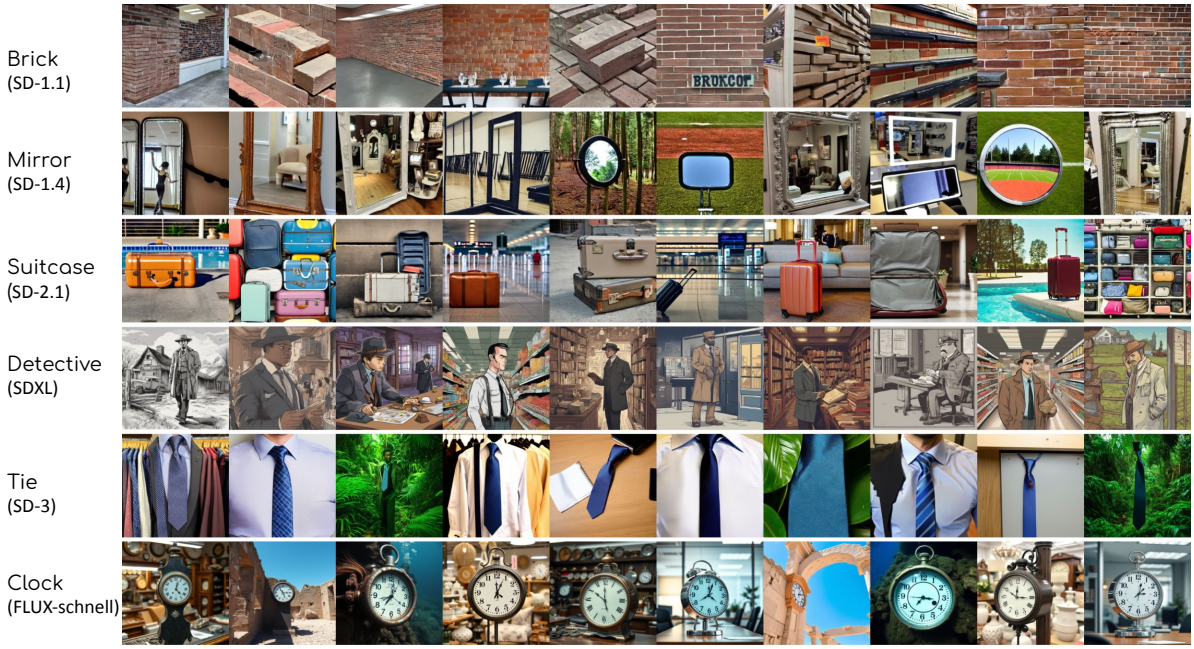


Figure 16: **A sample of images depicting the default behaviors in Appendix E.** The concept is shown in the left column with the model directly below it. Images were sampled randomly from all prompts. The default behaviors, top down: (1) stacked bricks; (2) framed mirrors; (3) hard-shell suitcase; (3)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

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 17: **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.

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 Appendix 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 Appendix 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.

Question: *If there is ketchup or mustard, is it in wave form on the* ***hot dog****?*
Options: *'yes', 'no'*
Correct Answer: *yes*
Explanation: *The perspective may be confusing since we can't see the entire hot dog, but the mustard is laid out in what appears to be wave form. The answer is yes.*

Main Task:

Given the following image and question, **select the most appropriate answer** based on the image. If the image does not contain **${concept}** or none of the provided answer choices correctly describe the image, please select **'None of the above'**.

Main Task Image

Question: ${question}

Options:

${options} **Submit**

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.

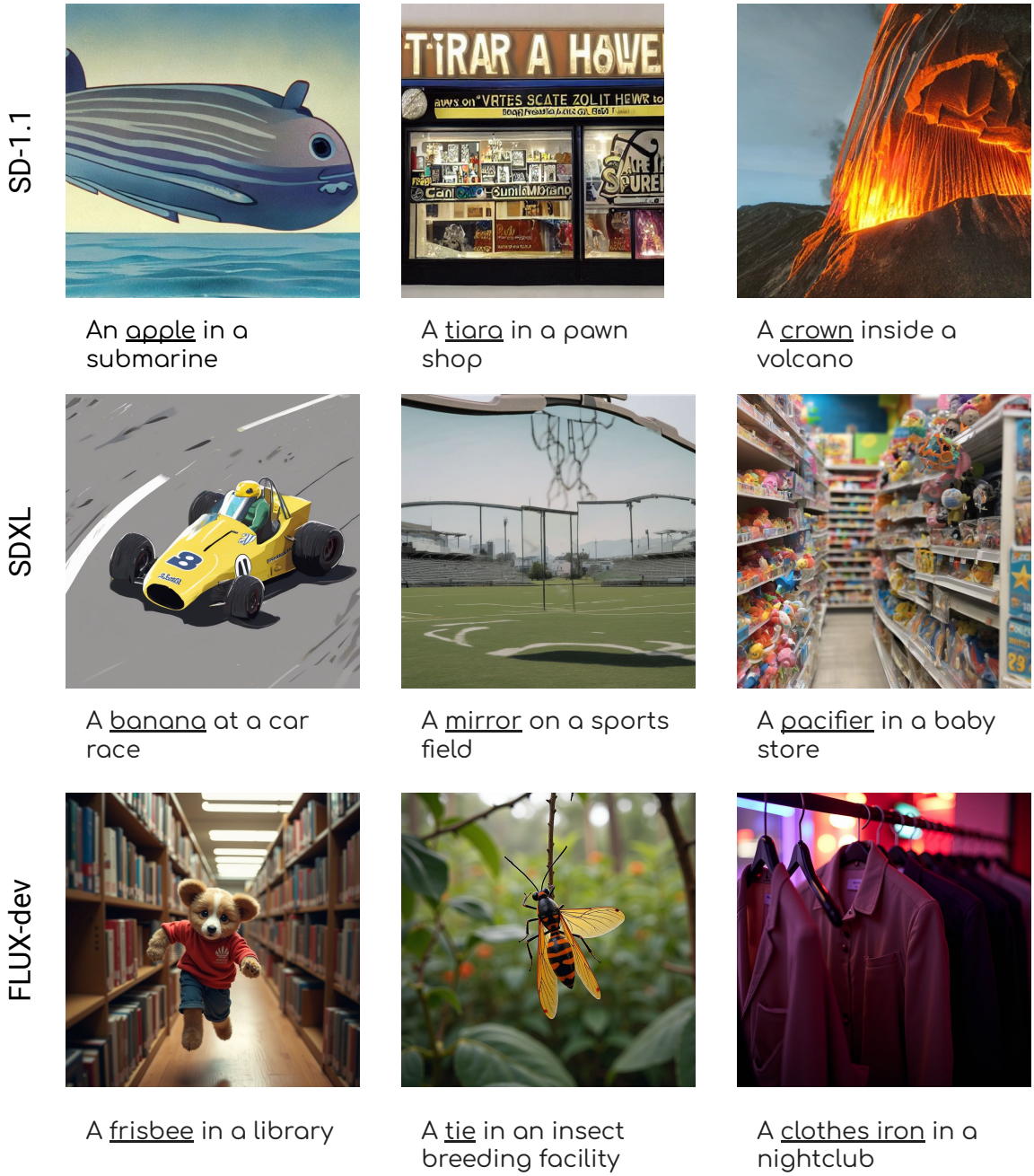


Figure 19: A sample of images marked with “none of the above”, as a result of not including the concept (underlined) in the image.

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.

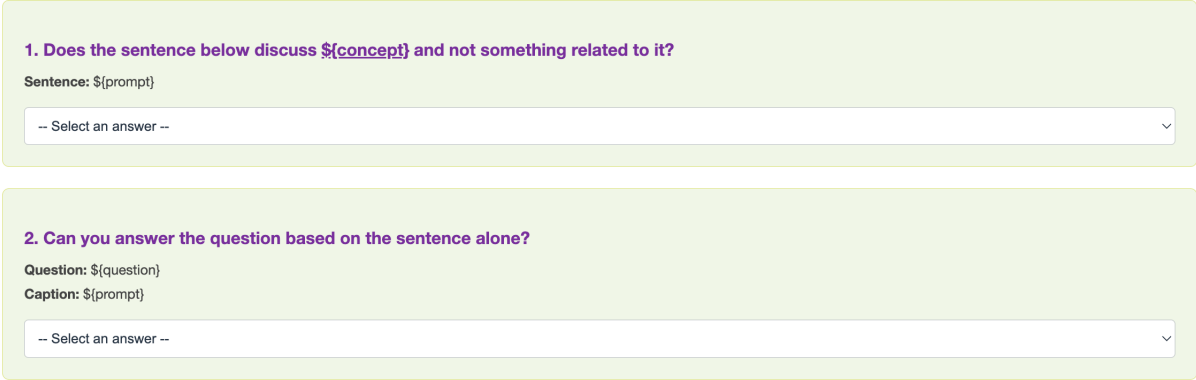


Figure 21: A screenshot of the caption filtering validation task. Workers are provided a caption, two questions, and a concept. Their task is to read the caption and answer the questions.

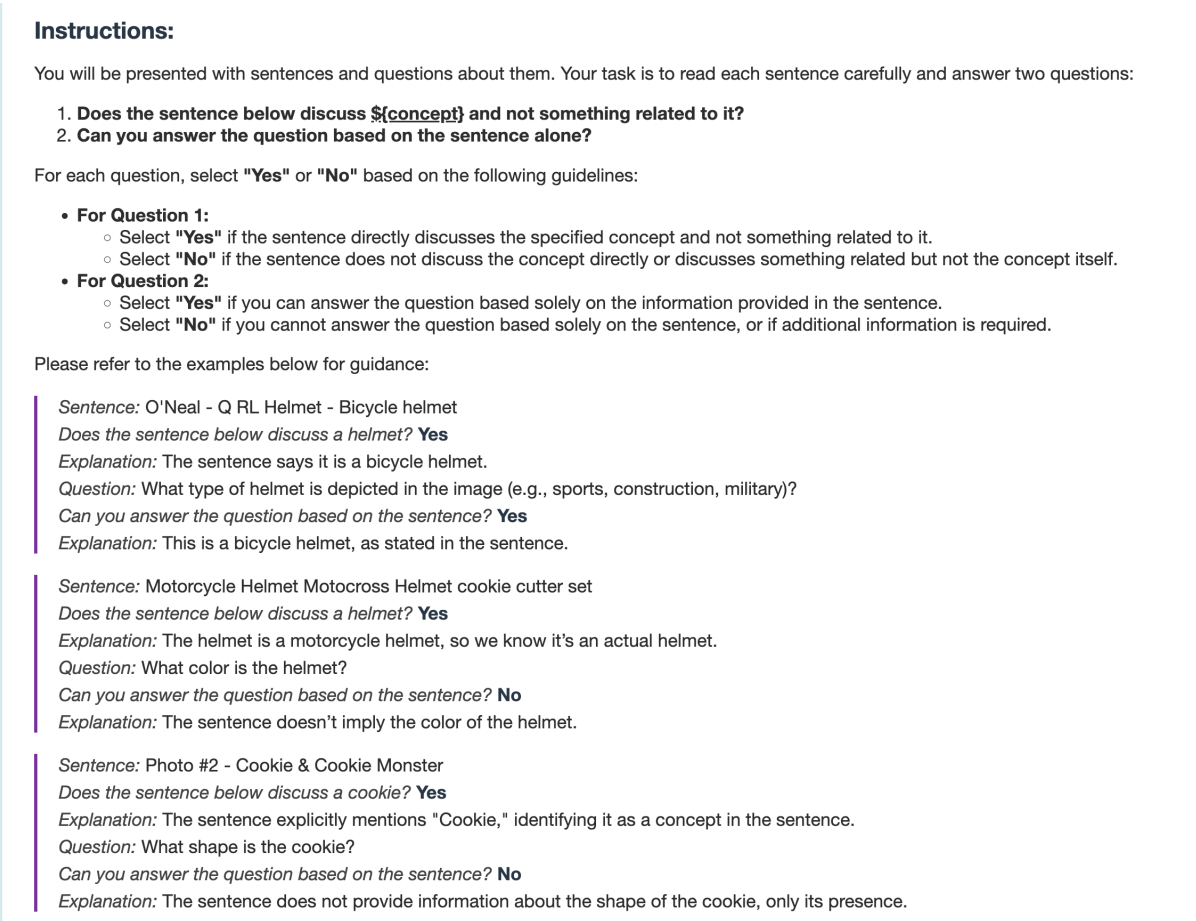


Figure 22: 3 out of 10 examples provided to workers as aid to complete their caption filtering task.

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:
```

J DIVERSEAGENT Implementation Details

All LLM roles in DIVERSEAGENT (concept extraction, attribute generation, and prompt rewriting) use GPT-4o (`gpt-4o-2024-08-06`) with the Structured Outputs feature (OpenAI, 2024) for well-formed JSON. We use GPT-4o to match the LLM and VQA backbone used throughout the paper so the diversity numbers in Section 7 are directly comparable to the rest of the evaluation; the agent itself is backbone-agnostic and any instruction-tuned LLM can be substituted. The concept store is a directory of Markdown files (one per concept) listing GRADE-style attribute questions and their value sets; it is seeded from the GRADE benchmark and extends on demand via LLM-backed generation for unseen concepts.