Benchmarking Diversity in Text-to-Image Models via Attribute-Conditional Human Evaluation

Isabela Albuquerque Google DeepMind

Ira Ktena Google DeepMind

Olivia Wiles Google DeepMind

Ivana Kajić Google DeepMind

Amal Rannen-Triki Google DeepMind

Cristina Vasconcelos Google DeepMind

Aida Nematzadeh Google DeepMind

Abstract

Despite advancements in photorealistic image generation, current text-to-image (T2I) models often lack diversity, generating homogeneous outputs. This work introduces a framework to address the need for robust diversity evaluation in T2I models. Our framework systematically assesses diversity by evaluating individual concepts and their relevant factors of variation. Key contributions include: (1) a novel human evaluation template for nuanced diversity assessment; (2) a curated prompt set covering diverse concepts with their identified factors of variation (e.g. prompt: AN IMAGE OF AN APPLE, factor of variation: color); and (3) a methodology for comparing models in terms of human annotations via binomial tests. Furthermore, we rigorously compare various image embeddings for diversity measurement. Our principled approach enables ranking of T2I models by diversity, identifying categories where they particularly struggle. This research offers a robust methodology and insights, paving the way for improvements in T2I model diversity and metric development.

1 Measuring diversity in text-to-image models

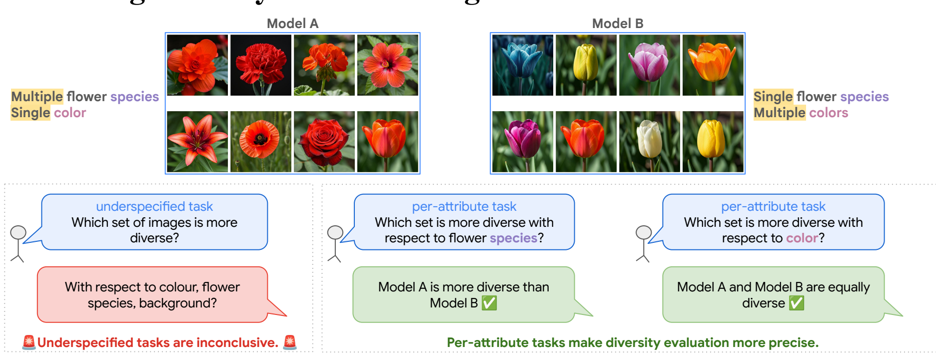


Figure 1: Evaluating diversity requires specifying both the concept being assessed and the factor of variation to reduce ambiguity in the annotation process.

Output diversity is widely considered desirable for text-to-image (T2I) generation models aiming to accurately represent the natural variability of entities in the real world. This is crucial not only technically, for serving as faithful world models, but also for downstream applications like supporting creative processes and ensuring broad conceptual representation across contexts. For example, a diverse model generating "an image of a house" should produce variations in architectural style and background. However, current diversity metrics often conflate it with other properties like fidelity

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

(e.g., Fréchet Inception Distance (FID) [14]). While progress has been made by developing dedicated metrics (e.g., Vendi Score [10]), the conditions for measuring diversity remain poorly defined and lack standardization, highlighting the need for a principled framework.

In particular, previous work often measures the variability of generated images in scenarios that do not explicitly account for diversity. For instance, images may be generated using a prompt set that neither requires nor controls for output variations [e.g., 31, 1], or models may be compared using a generic human evaluation template that does not specifically probe for diversity [e.g., 3]. This can result in measures of diversity that are ambiguous or inconclusive (see Fig. 1). To address this challenge, we propose a framework to measure diversity without conflating constructs [43, 44, 25, 16, 41]: we operate under the premise that systematically evaluating diversity requires specifying both the concept being assessed and the attribute of interest, as illustrated in Fig.1. We empirically validate this by demonstrating that human accuracy in evaluating diversity is at chance level when the attribute is not defined. Building on this observation, we introduce a novel evaluation framework designed to measure the per-attribute intrinsic diversity of T2I models. This framework includes a synthetically generated prompt set spanning common concepts and their variations, as well as a human evaluation template. The template, informed by empirical findings on a golden set, improves human accuracy by dividing the evaluation into two subtasks: counting and counts comparison.

Considering the high cost of human evaluations for model ranking, developing automated metrics that accurately reflect human judgment is crucial for advancing T2I models. While various diversity metrics have been proposed [10, 16], their alignment with human perceptions of diversity often remains unevaluated. To address this, we use our proposed human evaluation template and prompt set to examine the reliability of autoevaluation metrics. Specifically, we investigate the Vendi Score [10], a widely adopted diversity metric [19, 12] whose correlation with human-perceived diversity has not yet been thoroughly established. Our analysis reveals that the Vendi Score, when optimized for the appropriate representation space, can achieve approximately 65% accuracy in capturing human diversity judgments. We also find that the accuracy improves to 80% when the model pairs are more different, highlighting the need for more discriminant representations. Furthermore, we apply our framework to compare five recent generative models: Imagen 3 [2], Imagen 2.5 [39], Muse 2.2 [5], DALLE3 [3], and Flux 1.1 [20]. This comparison identifies Imagen 3 and Flux 1.1 as the top-performing models regarding attribute diversity. We believe our framework provides a robust foundation for future work in developing more human-aligned evaluation metrics and improving T2I model diversity. This research makes three key contributions:

- It formalizes the problem of quantifying diversity in T2I models and proposes a practical evaluation approach using pre-defined factors of variation.
- It introduces an evaluation framework consisting of a detailed prompt set (covering 86 concept-factor variation pairs) and a validated human evaluation template.
- It applies this framework to evaluate prominent T2I models and automatic evaluation metrics.

2 The three ingredients for diversity evaluation

To evaluate diversity, our framework is based on three components: a definition of what specific diversity is being measured, a prompt set to elicit relevant outputs, and a human evaluation template for reliably comparing models. These are described below.

2.1 A clearly specified problem: Diversity per attribute

Prelude: formalizing diversity. Consider a set of images $X = \{x_1, x_2, \dots, x_n\}$, where each image x_i belongs to a space $\mathcal{X} \subseteq \mathbb{R}^D$. We posit that the visual appearance of each image x_i is primarily determined by a set of K underlying independent generative factors $f_i = \{f_i^1, \dots, f_i^K\}$. A potential generative model could be formulated as:

$$p(x_i) = \prod_{k=1}^{K} p(x_i|f_i^k)p(f_i^k). \quad (1)$$

We focus on scenarios where images represent scenes containing instances from well-defined concepts (e.g., bottle, forest). Given a concept, we can often map these abstract generative factors to concrete, observable attributes. For instance, an image x_i depicting a bottle can be described by attributes such as: $f^{\text{material}} \in \{\text{glass}, \text{plastic}, \text{metal}\}$, $f^{\text{shape}} \in \{\text{cylindrical}, \text{square}\}$, and $f^{\text{state}} \in \{\text{open}, \text{closed}\}$.

Let $C = \{c^1, \dots, c^J\}$ be the set of concepts, $A^j = \{a^{j,1}, \dots, a^{j,K}\}$ the relevant attributes for a given concept c^j, and $V^{j,k}$ the finite set of possible values for attribute $a^{j,k}$. Each image x_i depicting a concept is associated with a specific value $v_i^{j,k} \in V^{j,k}$ for each attribute $a^{j,k}$. We define a sample of images X^j (for the same concept c^j) as *perfectly diverse* if it comprehensively covers all attribute variations. More precisely, for every attribute $a^{j,k} \in A^j$ and every possible value $v \in V^{j,k}$ there must exist at least one image $x_i^j \in X^j$ such that the attribute $a^{j,k}$ for image x_i^j takes the value v.

A tractable notion of diversity. Measuring diversity across the complete set of generative factors underlying natural data is significantly challenging. Firstly, the sheer number of potential factors (K) is often immense. Secondly, as highlighted by Tsirigotis et al. [38], the combination of their possible values grows exponentially, leading to a 'curse of generative dimensionality' where no realistic finite sample can cover all possible combinations. Thirdly, many factors may inherently possess continuous value ranges, making exhaustive coverage impossible even for a single factor.

Given these challenges, and since achieving the *perfect diversity* (as defined earlier) is intractable with a finite sample, we instead propose to measure *tractable diversity*. This approach focuses on a carefully selected subset of the most salient and practically relevant generative factors (K') for a specific concept. Identifying which factors are practically relevant is non-trivial and must be tailored for a given use case. In this work, to identify these factors, we focus on commonly observed concepts reflective of T2I model training data. To effectively sample from the distribution of generative factors within these concepts, we leverage the knowledge encoded by Large Language Models (LLMs) [30]. Specifically, we prompt an LLM (Gemini 1.5 M [37]) to identify relevant aspects of variation for evaluating the diversity of a given concept. The full system instruction is given in the Appendix.

2.2 A systematically generated prompt set

Our goal is to rigorously evaluate generative models and diversity metrics, specifically focusing on their ability to represent variation within distinct attributes of concepts. To effectively rank these models and metrics, our evaluation framework must accommodate both precisely controlled scenarios and complex, real-world use cases. We deliberately select concepts that are ubiquitous in everyday life and common image datasets, such as ImageNet [8] (e.g., 'fruit', 'car', 'snake'), thereby anchoring our evaluation in practical utility. However, simple concepts alone are insufficient. They must also possess inherent complexity and variability, presenting a genuine challenge to the models and metrics. The chosen concepts and their attributes need to be sufficiently nuanced to allow our evaluation methodology to clearly reveal performance differences and track improvements over time or across different systems.

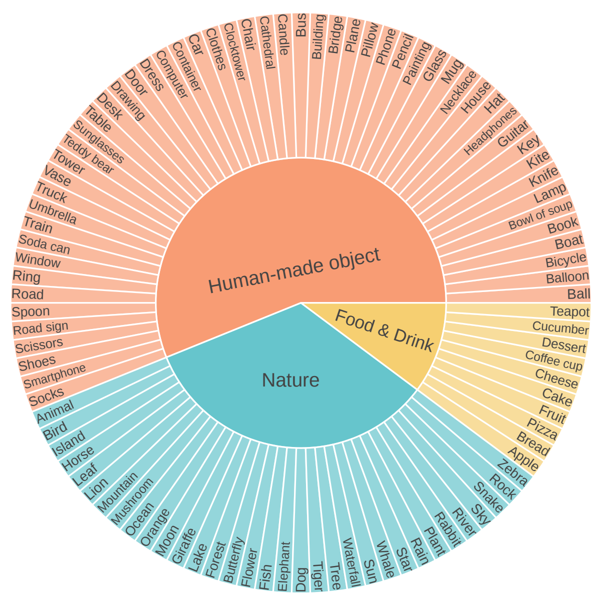


Figure 2: Each slice represents a concept, grouped and color-coded by its overall category.

To structure this process, we classify concepts into three widely applicable categories: *Food and Drink* (items like *coffee cup*, *cake*), *Nature* (elements like *river*, *butterfly*), and *Human-made Objects* (artefacts like *bridge*, *laptop*). We leverage the generative capabilities of Large Language Models (LLMs) to systematically produce a wide range of concepts within these categories. The instruction to generate "ImageNet-like" concepts guides the LLM towards producing concrete, typically visualizable nouns, similar in scope to those in large-scale image datasets. For each generated concept, we then perform a subsequent step, again using an LLM, to identify a semantically relevant aspect of variation (attribute) that is intrinsic or commonly associated with that concept. This yields concept-attribute pairs $(c^j, a^{j,k})$ such as: (*apple*, *color*), (*car*, *type*), (*tree*, *species*), (*coffee cup*, *material*), (*chair*, *style*). This two-stage, LLM-driven process allows us to systematically build a prompt set specifically designed to probe and evaluate diversity along meaningful, contextually relevant dimensions for a broad range of common concepts. Finally, the authors manually verified all concept-attribute pairs and removed 5 where the attribute was potentially difficult / ambiguous to categorize (e.g. (*food*, *cuisine*)).

2.3 A validated, bespoke human evaluation template

Prior work has shown that developing an appropriate human evaluation template is an essential component in the process of measuring a desired capability of a generative model [42, 7]. To that end, we develop a human evaluation template that: (a) allows annotators to understand the task well, (b) captures their judgment faithfully, and (c) yields meaningful ground truth annotations for per-attribute diversity, subsequently used to validate automated evaluation metrics. The annotators are provided with 4 options for the side-by-side comparison: (i) Left more diverse, (ii) Right more diverse, (iii) Equally diverse, (iv) Unable to answer.

A template to measure per-attribute diversity. Our template for measuring per-attribute diversity employs a comparative, side-by-side approach due to the difficulty of evaluating diversity within a single set. Many existing diversity metrics also require a reference set. We considered the following design choices for our human evaluation template to ensure meaningful assessment (1) *Set size*: Balancing the perception of diversity with minimizing annotation fatigue and enabling robust computation for metrics requiring larger sets (e.g., Vendi score). (2) *Attribute specification*: Explicitly stating the attribute for evaluation versus allowing open-ended diversity assessment. (3) *Anchoring task*: Incorporating an intermediate task to guide annotators to focus on the intended attribute.

Validating the template with a golden set. To evaluate the quality of the evaluation template, we curate a golden set of 10 <concept, aspect> pairs, where concept corresponds to a concept that should be considered common across images in a set and aspect describes the associated aspect of variation that we want to measure diversity against. We validate the evaluation template by comparing cases where (i) the concept *remains constant* across images in the set while the aspect *varies* (ii) the concept *varies* across images while the aspect *remains the same*, and (iii) *both* the concept and the aspect vary across images within the set. We expect images in set (i) to be considered more diverse than images in set (ii), and similarly images in set (iii) to be considered more diverse than images in set (ii). Finally, we expect that images in sets (ii) and (iii) are considered equally diverse as we want to focus on the aspect as axis of variation.

In Fig. 3, we present the annotation accuracy of human experts using our template under various conditions, treating our definitions (in the previous paragraph) as ground truth. The different templates are shown in Fig. 9. The accuracy for the w/o aspect task is 30.0% for comparisons of sets of size 4 and 26.7% for sets of size 8. In contrast, the template that includes the aspect shows a significant increase in accuracy (82.5% for set size 4 and 53.3% for set size 8), indicating that explicitly mentioning the desired aspect of variation improves annotation accuracy. This improvement likely stems from preventing annotators from unintentionally conflating the concept and the aspect when not guided to focus on a specific axis. Furthermore, we observe that adding the count anchoring question enhances accuracy, especially for the set size of 8, reaching 77.9%.

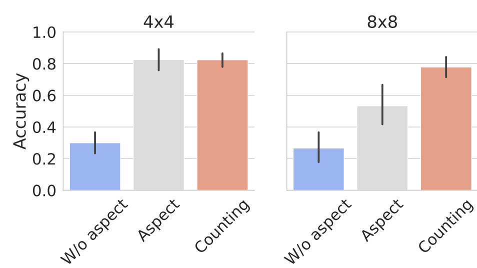
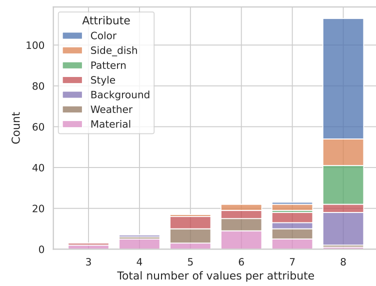
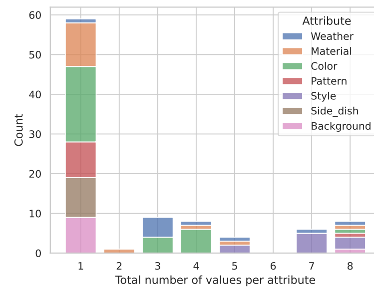


Figure 3: Match with the golden set depending on different set sizes.

For the count task, we found a strong ($\rho = 0.88$) and statistically significant ($p < .001$) correlation between the annotators' final diversity comparison and the comparison inferred from their individual subset counts (where a higher count on one side implies a more diverse final response for that side, and equal counts imply equal diversity). This confirms that the anchoring count question effectively guides annotators. To further validate our setup, we analyzed instances where annotators' responses deviated from the ground truth in our golden set. We examined the distributions of attribute counts for two image subsets: (1) those labelled "diverse" in the ground truth, where we expected a count mode of "8" and (2) those labelled "non-diverse", where we expected a mode of "1". The



(a) The “diverse” golden set.



(b) The “non-diverse” golden set.

Figure 4: The distribution of counts for sets of images labelled as “diverse” or “non-diverse” in the golden set for the pilot study.

results of this analysis are presented in Fig. 4. While generally, annotator responses aligned with the golden set labels, we observed a few exceptions. For instance, in one case labelled as a diverse set of chairs, all annotators counted only 3 or 4 distinct chair types, indicating lower diversity than expected. Upon closer inspection, these chairs appeared visually similar despite potentially different underlying material prompts (e.g., metal, iron, aluminum).

3 Our framework in practice

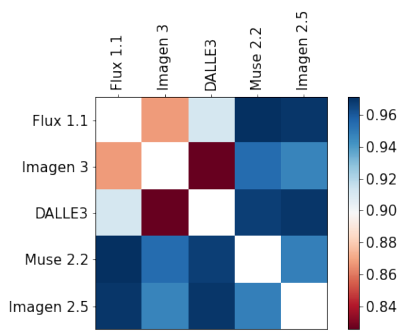
We demonstrate our framework’s practical application by: (i) collecting comprehensive human annotations with our template to compare models, (ii) using these annotations as ground truth to evaluate diversity metrics, and (iii) comparing model rankings from human versus automatic evaluations to highlight the gap between human-perceived diversity and current metric capabilities.

3.1 Ranking models via human evaluation

With the proposed prompt set from Sec. 2.2 and the human evaluation template introduced in Sec. 2.3, we evaluate the attribute-based diversity of five generative models, namely: Muse 2.2 [5], Imagen 2.5 [39], Imagen 3 [2], DALLE3 [3], and Flux 1.1 [20]. For each model, we generate 20 distinct samples for each prompt, randomly combine them in 10 different sets of 8 images, and run side-by-side evaluations for all 10 combinations of 2 models. For each side-by-side comparison, evaluations from 5 different raters were collected. Raters had access to a slide deck with instructions to perform the task and were compensated for the time invested in the data collection. Details can be found in the Appendix (Sec.B) Before comparing each model pair in terms of diversity, we evaluate the overall annotations quality by computing the inter-annotator agreement via Krippendorff’s alpha reliability (α) [11]. In Fig. 5a, we observe that for all cases $\alpha > 0.8$, indicating a high-degree of agreement across annotators [24].

Ratings aggregation. Given the high levels of inter-annotator agreement for all runs of the human evaluation, we aggregate annotations for each side-by-side comparison across raters by *taking the mode* of the ratings. We then follow this step with a second aggregation, this time at the level of all side-by-side comparisons for each concept. For instance, when comparing a given model pair, there are 10 side-by-side comparisons for the concept *apple* (each side-by-side comparison here corresponds to the evaluation of two sets of 8 images). At the end of this process, for the considered models pair, we obtain a single human evaluation result for each concept in the prompt set.

Model ranking. Using the results from the ratings aggregation, we propose to use Binomial tests to verify the following hypothesis: *there is a significant difference between the outcomes of a given pair of models*. To do so, we count the number of categories for which each model was deemed best and perform a two-sided Binomial test under the null-hypothesis that the rate for which each model is the best for a concept is equal to 50% (i.e. both models have equal win rates). Results considering a 95% confidence level for all tests are shown is Fig. 5b. Imagen 3 and Flux 1.1 are significantly better or not worse than all other models. Imagen 2.5 and Muse 2.2 are not significantly better than any contender, showing that our benchmark is able to capture an overall progress in diversity when comparing newer and older models. DALLE3 is significantly better than Imagen 2.5, but does not significantly surpass the performance of the other models considered for comparison.



(a) Krippendorff's α-reliability.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	=	>	>	>
Imagen 3	=	×	=	>	>
DALLE3	<	=	×	=	>
Muse 2.2	<	<	=	×	=
Imagen 2.5	<	<	<	=	×

(b) Binomial test results at a 95% confidence level.

Figure 5: **Human evaluation results.** (a) Inter-annotator agreement results in terms of Krippendorff's α-reliability. (b) We compare model rankings in terms of significance in the number of wins with two-sided Binomial tests under a 95% confidence level. Each entry in the grid represents a comparison between two models. The sign indicates the model in the row is better ($>$), worse ($<$), or not significantly different (=) than the model in the column.

3.2 Comparing autoevaluation metrics

While human evaluation is often considered gold standard, it can be impractical to rely solely on human annotation. We then leverage the collected human annotations to perform an extensive study of the role of embeddings for the Vendi Score [1].

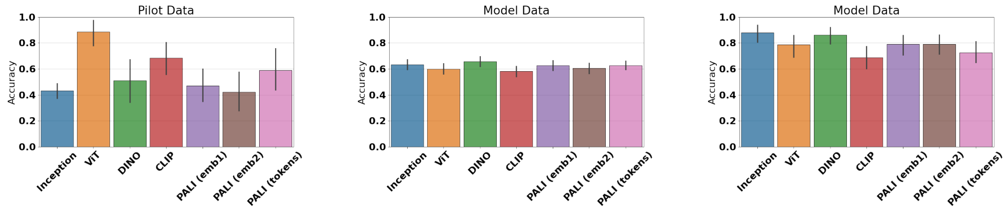
Autoraters based on the Vendi Score. Given a set of images $X^{j,k} = \{x_i^{j,k}\}$ (corresponding to a given model, concept c^j and attribute $a^{j,k} \in A^j$), we extract embeddings $h_\Xi(x_i^{j,k})$ for each image. h_Ξ is a pretrained feature extractor that can be dependent on a set of conditions $\Xi = \{\xi_l\} \subset (C \times A) \cup \{\xi^0\}$ where ξ^0 is a condition unrelated to the considered categories and attributes that can be added to test the impact of conditioning. The different feature extractors and conditions we used are detailed in the following paragraph, but here are a few generic examples to clarify the notation: (i) h_Ξ takes only images as input. In this case, $\Xi = \emptyset$. (ii) h_Ξ is a vision and language model. In this case, embeddings can be conditioned on text data that depends on the concept only (i.e., $\Xi = \{c^j\}$), attribute only (i.e., $\Xi = \{a^{j,k}\}$), or both concept and attribute (i.e., $\Xi = \{c^j, a^{j,k}\}$). To test the impact of conditioning on text, we can instead choose an unrelated prompt (i.e., using $\Xi = \{\xi^0\}$). Finally, we aggregate the embeddings using a diversity metric to obtain a score for the set. As we do not have access to a reliable reference in our setting, we use the Vendi Score [10], a reference-free and widely adopted metric [28, 16, 12, 18]. The Vendi Score is defined as follows:

Definition 1 (Adapted from [10], Definition 3.1)**.** *Given a concept c^j, an attribute $a^{j,k}$ and a set of conditions Ξ, let $\{x_1^{j,k}, \ldots, x_n^{j,k}\}$ denote a set of images representing a given concept and attribute. Let $k : X \times X \to \mathbb{R}$ be the cosine similarity between the embeddings of two images, $K^\Xi \in \mathbb{R}^{n \times n}$ be the kernel matrix, with $K^\Xi_{lm} = k^\Xi(x_l^{j,k}, x_m^{j,k})$, and let $\lambda_1^\Xi, \ldots, \lambda_n^\Xi$ be the eigenvalues of K^Ξ/n. The Vendi Score for the set $\{x_1^{j,k}, \ldots, x_n^{j,k}\}$ is defined as:*

$$s_\Xi(x_1^{j,k}, \ldots, x_n^{j,k}) = \exp(-\sum_{i=1}^{n} \lambda_i^\Xi \log \lambda_i^\Xi). \quad (2)$$

Experimental setup. We compare three different types of embeddings. First, we compare embeddings obtained *using only* the image input. Here we consider two models trained for IMAGENET classification – the IMAGENET INCEPTION model introduced in [36] and an IMAGENET VIT-B/16 model trained on IMAGENET21K as described in [35]. We also consider one self-supervised model, DINOV2 [26]. Second, we consider embeddings conditioned on both the image and textual attribute. We use PALI embeddings [4] at various points after fusing the text and visual input, and CLIP [29] combined text and image embedding. We use these embedding models to obtain an embedding for

[1] Results with other autoraters can be found in the Appendix Sec.**??**.



(a) The “diverse” golden set.

(b) Side-by-side model comparisons.

(c) Side-by-side model comparisons with diversity gap > 4.

Figure 6: **Autoevaluation results:** the performance of the Vendi Score given different embeddings across three settings: (a) the golden set; (b) all the annotations gathered; (c) the “easy” subset of the annotations where raters identified a diversity gap of > 4 for a pair. On the golden set, VIT performs best but this does not transfer to side-by-side comparisons. The performance is generally better on the “easy” split of the data, showing that the embeddings perform considerably worse when the difference between the generated sets of images is more subtle—models are more similar.

each image in a set. We then use the Vendi Score in order to aggregate embeddings and obtain a diversity prediction for the set. Finally, we consider the first word output by the PALI model as a discrete token. We aggregate these outputs by counting the number of unique words generated for a set to get an estimate for diversity.

For each pair of image sets, we analyze the agreement between a diversity assessment based on our autoraters, and the assessment resulting from the human annotations, not taking into account pairs where the annotators found the sets to be equally diverse. If the autoraters and the human evaluations both indicate the same set as being the most diverse (i.e., $s_\Xi(X_1^{j,k}) > s_\Xi(X_2^{j,k})$ and annotators rated the set $X_1^{j,k}$ generated with model 1 based on concept c^j and attribute $a^{j,k}$ as more diverse than $X_2^{j,k}$ generated with model 2 based on the same concept and attribute), we say that for that pair of sets, the autorater is correct, else it is incorrect. We then report accuracy by aggregating the number of pairs for which the autoraters are correct. Results are reported in Figs. 6a-6c. We can see that, on the “diverse” golden set, the VIT model does the best, and then the tokens of PALI. This is perhaps surprising, as the VIT model is not specifically trained to focus on the aspects we are considering for diversity but to be able to discriminate between broad classes. However, we see minimal difference in results if we consider the model data. All approaches perform similarly and lead to accuracies that are not significantly different. We hypothesize that the reason for the observed small difference in results was that the models were similar to each other. As a result, we looked at ratings where the annotators perceived a larger gap between models by using the counts as a proxy. We consider a subset of the data where the difference in counts between the two sets is greater than 4, keeping about 24% of the data. We find that now, on the model data we see a bigger difference in results. First, all autoraters are more accurate. Second, we can see that again the image based approaches (e.g., the INCEPTION model, the DINO model and VIT model) perform best.

In Figs. 7 and 15 we visualize examples for four side-by-side comparisons where the corresponding autoraters indicate that a group of images have highest or lowest diversity. We can see that results are reasonable and that in general, images with low diversity arise due to mode collapse, i.e. the model generates a very similar image for the same concept. This could explain why the INCEPTION model performs poorly on the pilot data but well on the model comparison data. INCEPTION features are effective for identifying these issues but no effective for identifying diversity in the case of confounding aspects (e.g., the background is changing while the animal is staying the same).

3.3 Ranking models with autoevaluation approaches

Ranking is achieved by counting the frequency at which the model on the left achieves a higher score than the model on the top (i.e. for "model 1" on the left axis and "model 2" on the top, we count how many times $s_\Xi(X_1^{j,k}) > s_\Xi(X_2^{j,k})$, with $X_1^{j,k}$ generated with model 1, and $X_2^{j,k}$ generated with model 2), and subtracting 0.5. The win rate matrices with all models and the score distributions for Imagen 3 and Flux 1.1, the two models that were preferred by human annotators, are shown in Sec. D.5 in the Appendix. In order to test the significance of these comparisons, we aggregate the scores per concept and perform a Wilcoxon signed-rank test under a 95% confidence

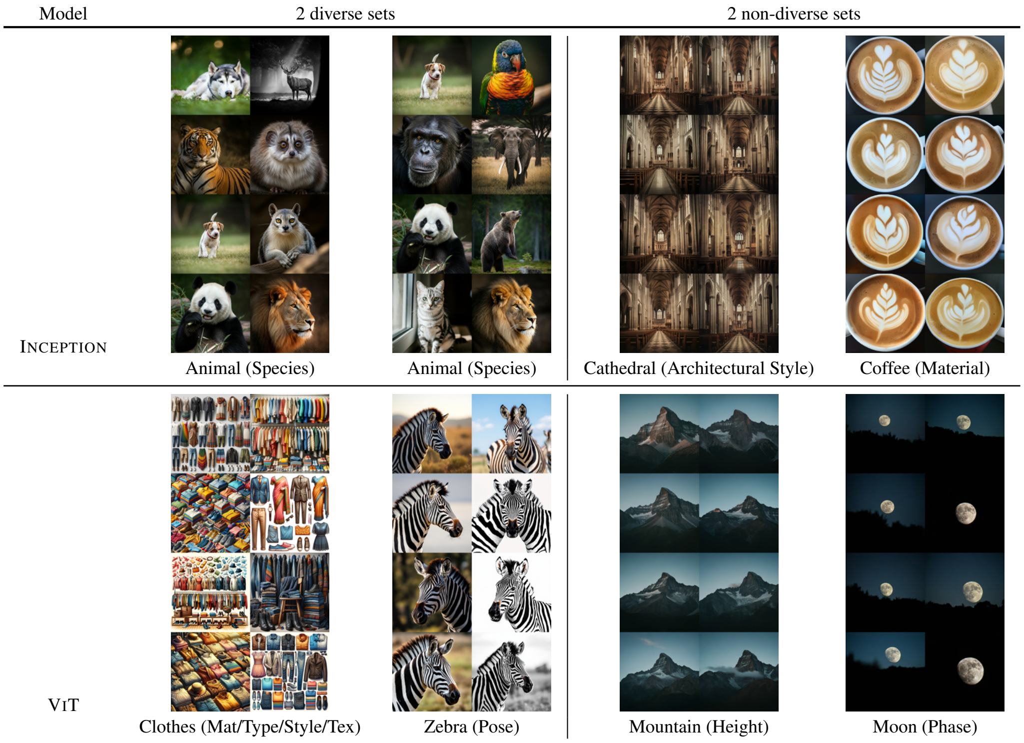


Figure 7: Qualitative results for different autoraters on the T2I annotated dataset, showing two very diverse and two non diverse sets as determined by the autorater.

level. On the left panel, we consider the IMAGENET INCEPTION embeddings, as they yielded the highest accuracy on the model data. In the middle and the right panels, we consider text-conditioned embeddings, as they are closest to our human evaluation procedure. We show the results using PALI(EMB1), as they show a marginal advantage on model data. On the middle panel, we show the results corresponding to conditioning the embedding model on the attribute only, while on the right panel, conditioning takes into account both attribute and object. Results with other embeddings can be found in the Appendix (Sec D.5). Through the autoevaluation model ranking, we find that independently of the chosen embedding, Imagen 3 is not worse than all other models, and Flux 1.1, Imagen 3 and DALLE3 are better than Imagen 2.5 and Muse 2.2. We also observe that using the IMAGENET INCEPTION embeddings and the PALI(EMB1) with a conditioning on object and attribute captures more differences across the three top models, and that using both types of the PALI(EMB1) embeddings captures more differences between Imagen 2.5 and Muse 2.2.

By adopting the model comparison results obtained with the human annotations as shown Fig. 5b as ground-truth, we find that all used embeddings are of similar quality in terms of closeness to human perception of diversity. They all did not flip conclusions, but the autoevaluation approach seems more sensitive to certain variations depending on the choice of embedding model and conditioning. Text conditioning, while closest to the human evaluation procedure, did not show a significant advantage with the current choice of embedding models and conditioning. However, we observe in Fig. 8 the influence of the conditioning. The additional results in the Appendix (Sec. D.5) show the influence of the choice of embedding models. It is possible that better choices of models and conditioning prompts can lead to better results, but we leave this question open for future investigation.

4 Related work

The primary method for evaluating text-to-image models involves gathering human judgments on a specific benchmark (i.e., a set of prompts). Previous research highlights that the composition of this benchmark significantly influences the resulting model rankings. This has led to the development of benchmarks with broader skill coverage, e.g., text rendering and spatial reasoning [6, 21, 42], as well as benchmarks targeting specific skills like numerical reasoning [17]. Although human evaluation remains the gold standard, numerous automatic metrics have been proposed to potentially

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	<	=	>	>
Imagen 3	>	×	>	>	>
DALLE3	=	<	×	>	>
Muse 2.2	<	<	<	×	=
Imagen 2.5	<	<	<	=	×

(a) Inception embeddings.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	=	=	>	>
Imagen 3	=	×	=	>	>
DALLE3	=	=	×	>	>
Muse 2.2	<	<	<	×	>
Imagen 2.5	<	<	<	<	×

(b) PALI(emb1) embeddings - conditioned on attribute.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	<	=	>	>
Imagen 3	>	×	>	>	>
DALLE3	=	<	×	>	>
Muse 2.2	<	<	<	×	>
Imagen 2.5	<	<	<	<	×

(c) PALI(emb1) embeddings - conditioned on object and attribute.

Figure 8: **Ranking by autoevaluation.** We compare model pairs given the Vendi Score based on (a) Inception, (b) PALI(emb1) conditioned on the attribute, and (c) PALI(emb1) conditioned on object and attribute. Each entry in a grid represents a comparison between two models. Significance is tested via the Wilcoxon signed-rank under a 95% confidence level. The sign indicates the model in the row is better (>), worse (<), or not significantly different (=) than the model in the column.

replace human judgments, at least for certain applications [e.g., 13, 42, 15, 23, 34]. Rigorous validation of these metrics is crucial across diverse conditions, including different prompt sets, human evaluation templates, and models [42]. An important facet of evaluating text-to-image models involves measuring the diversity of their output [9, 40]. This has resulted in different metrics, both reference-based [32, 14, 33] and reference-free [10, 30, 25, 27, 22]. The advantage of reference-free metrics is their independence from a ground-truth set, which permits the evaluation of diversity in broader contexts. One such recent metric, the Vendi score [10], has influenced subsequent research [18, 12, 16]. Despite these developments, none of the proposed metrics have undergone thorough evaluation, frequently being tested only on generic prompts or in simplified settings. Moreover, surprisingly, the majority of previous studies lack human evaluation to demonstrate the validity of these metrics. To address this gap, we introduce a prompt set designed for evaluating diversity across particular attributes and propose and validate a human evaluation template to gather ground-truth diversity judgments. Finally, we compare existing metrics and models under various conditions.

5 Discussion

Ensuring diversity in text-to-image (T2I) model outputs is essential, serving as a measure of their ability to express real-world variety. However, rigorous evaluation of this diversity, particularly for specific attributes, remains challenging. This paper introduces a novel framework for attribute-specific T2I diversity evaluation. It comprises a systematic prompt set and a human evaluation template, which has been validated to significantly improve the accuracy of human judgments by explicitly defining the attribute of interest. This framework provides a crucial ground truth for understanding and measuring diversity beyond general impressions.

Applying this framework, we ranked prominent T2I models based on their attribute-specific diversity, identifying Imagen 3 and Flux 1.1 as strong performers. Furthermore, we leveraged our human data to evaluate automated evaluation approaches based on the Vendi Score. Our results demonstrate that the choice of embedding space, upon which autoevaluation metrics operate, is crucial for achieving results that broadly align with human judgments. Notably, our findings indicate that Vendi Score-based autoevaluation approaches can capture human-perceived diversity with approximately 80% accuracy and correctly yield similar results for pairwise model comparisons when a comparable statistical analysis methodology is employed. The proposed framework and our collected data are intended to encourage future work on both T2I model improvement and the development of more reliable evaluation metrics. The broad impact of this work lies in its potential to improve T2I model quality in terms of diversity by providing an evaluation framework grounded in human perception. Moreover, unlike the previous work that often relies on attribute classifiers (e.g., gender), our evaluation methodology can be employed to measure demographic diversity in a classification-free manner. This potentially contributes to the development of more responsible AI systems.

References

[1] P. Astolfi, M. Careil, M. Hall, O. Mañas, M. Muckley, J. Verbeek, A. R. Soriano, and M. Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint arXiv:2406.10429*, 2024.

[2] J. Baldridge, J. Bauer, M. Bhutani, N. Brichtova, A. Bunner, L. Castrejon, K. Chan, Y. Chen, S. Dieleman, Y. Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

[3] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[4] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[5] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[6] J. Cho, Y. Hu, R. Garg, P. Anderson, R. Krishna, J. Baldridge, M. Bansal, J. Pont-Tuset, and S. Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

[7] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] M. Dombrowski, W. Zhang, S. Cechnicka, H. Reynaud, and B. Kainz. Image generation diversity issues and how to tame them. *arXiv preprint arXiv:2411.16171*, 2024.

[10] D. Friedman and A. B. Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[11] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[12] R. A. Hemmat, M. Hall, A. Sun, C. Ross, M. Drozdzal, and A. Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. *arXiv preprint arXiv:2406.04551*, 2024.

[13] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[15] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[16] M. Jalali, A. Ospanov, A. Gohari, and F. Farnia. Conditional vendi score: An information-theoretic approach to diversity evaluation of prompt-based generative models. *arXiv preprint arXiv:2411.02817*, 2024.

[17] I. Kajić, O. Wiles, I. Albuquerque, M. Bauer, S. Wang, J. Pont-Tuset, and A. Nematzadeh. Evaluating numerical reasoning in text-to-image models. *Advances in Neural Information Processing Systems*, 37:42211–42224, 2024.

[18] N. Kannen, A. Ahmad, M. Andreetto, V. Prabhakaran, U. Prabhu, A. B. Dieng, P. Bhattacharyya, and S. Dave. Beyond aesthetics: Cultural competence in text-to-image models. *arXiv preprint arXiv:2407.06863*, 2024.

[19] N. Kannen, A. Ahmad, V. Prabhakaran, U. Prabhu, A. B. Dieng, P. Bhattacharyya, S. Dave, et al. Beyond aesthetics: Cultural competence in text-to-image models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[20] B. F. Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[21] B. Li, Z. Lin, D. Pathak, J. Li, Y. Fei, K. Wu, T. Ling, X. Xia, P. Zhang, G. Neubig, and D. Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. URL `https://arxiv.org/abs/2406.13743`.

[22] K. Limbeck, R. Andreeva, R. Sarkar, and B. Rieck. Metric space magnitude for evaluating the diversity of latent representations. *Advances in Neural Information Processing Systems*, 37: 123911–123953, 2024.

[23] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024.

[24] G. Marzi, M. Balzano, and D. Marchiori. K-alpha calculator–krippendorff's alpha calculator: a user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545, 2024.

[25] M. Mironov and L. Prokhorenkova. Measuring diversity: Axioms and challenges. *arXiv preprint arXiv:2410.14556*, 2024.

[26] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[27] A. Ospanov, J. Zhang, M. Jalali, X. Cao, A. Bogdanov, and F. Farnia. Towards a scalable reference-free evaluation of generative models. *Advances in Neural Information Processing Systems*, 37:120892–120927, 2025.

[28] A. P. Pasarkar and A. B. Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. *arXiv preprint arXiv:2310.12952*, 2023.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[30] R. Rassin, A. Slobodkin, S. Ravfogel, Y. Elazar, and Y. Goldberg. Grade: Quantifying sample diversity in text-to-image models. *arXiv preprint arXiv:2410.22592*, 2024.

[31] S. Sadat, J. Buhmann, D. Bradley, O. Hilliges, and R. M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024.

[32] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall, 2018. URL `https://arxiv.org/abs/1806.00035`.

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[34] N. K. Senthilkumar, A. Ahmad, M. Andreetto, V. Prabhakaran, U. Prabhu, A. B. Dieng, P. Bhattacharyya, and S. Dave. Beyond aesthetics: Cultural competence in text-to-image models. *Advances in Neural Information Processing Systems*, 37:13716–13747, 2024.

[35] A. P. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[37] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[38] C. Tsirigotis, J. Monteiro, P. Rodriguez, D. Vazquez, and A. C. Courville. Group robust classification without any group information. *Advances in Neural Information Processing Systems*, 36, 2024.

[39] C. N. Vasconcelos, A. Rashwan, A. Waters, T. Walker, K. Xu, J. Yan, R. Qian, Y. Li, S. LUO, Y. Onoe, et al. Greedy growing enables high-resolution pixel-based diffusion models. *Transactions on Machine Learning Research*, 2024.

[40] J. Vice, N. Akhtar, R. Hartley, and A. Mian. On the fairness, diversity and reliability of text-to-image generative models. *arXiv preprint arXiv:2411.13981*, 2024.

[41] S. Vrijenhoek, S. Daniil, J. Sandel, and L. Hollink. Diversity of what? on the different conceptualizations of diversity in recommender systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 573–584, 2024.

[42] O. Wiles, C. Zhang, I. Albuquerque, I. Kajić, S. Wang, E. Bugliarello, Y. Onoe, C. Knutsen, C. Rashtchian, J. Pont-Tuset, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*, 2024.

[43] D. Zhao, J. T. Andrews, O. Papakyriakopoulos, and A. Xiang. Position: Measure dataset diversity, don't just claim it. *arXiv preprint arXiv:2407.08188*, 2024.

[44] D. Zhao, J. T. Andrews, A. Sony, T. O. Papakyriakopoulos, and A. Xiang. Measuring diversity in datasets. In *International Conference on Learning Representations*, volume 1, page 36, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are regarding the dataset and human evaluation template are supported by Sections 2.2 and 2.3. Claims related to experimental results are supported by results in Sec. 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: For ensuring reproducibility of the human evaluation results, we describe in detail the procedure to collect human annotations in Sec. 2.3. For the auto-evaluation results, all details related to how the embeddings were extracted for all considered cases, as well as, hyperparameters for computing the Vendi Score are available in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset created in this work will be publicly available upon acceptance and is currently available to reviewers, ACs, and SACs as per the data submission policy for the Datasets and Benchmarks track.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental details in Sec. 3 as well as in the Appendix. Our work does not involve training models, but we do disclose the hyperparameters necessary to compute metrics.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide error bars and statistical significance tests for all major experiments in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Where possible (for models that were not call through an API), we stated the computational resources that were used to run the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the code of Ethics by, for example, ensuring participants in the data collection were fairly compensated by their time ($13.88 hourly wage) respecting the minimum hourly wage for their location.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We include a discussion on the broader societal impacts of our work in Sec. 5 and in the Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All text-to-image models used in our work are cited where appropriate and used in according to their respective lincenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduced a new set of prompts, with corresponding images and human evaluation results. The paper describe the process to collect the data which is available with the necessary documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We presented in the Appendix the instructions given to the annotations as well as details about the compensation.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: We included information regarding raters instructions (that included potential risks) in the Appendix.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We disclose the use of LLMs to generate our prompts in Sec. 2.2.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Supplementary Material

A Human evaluation task details

A.1 Instructions

Before completing the annotation task, annotators were given a comprehensive set of instructions including the following guidelines:

- The goal of the task is to compare the how diverse two sets of images are with respect to a given attribute;
- For the given two sets of images, answer the question about how diverse the concept is with respect to the specific attribute highlighted in the prompt;
- You should count how many different instances of a particular attribute they observe on the left and right sets of images, separately;
- For example, if the attribute is “background” and the prompt is “animal”, raters should count how many different backgrounds appear in each set of images and finally judge how diversity of the two sets compares to each other with respect to this attribute;
- Finally, based on the counts, pick one of the following options: (1) Left is more diverse; (2) Right is more diverse; (3) Equally diverse; (4) Unable to answer.

Along with the written instructions, annotators were also given examples corresponding to options 1, 2, and 3.

A.2 Additional information

In total, 24591 annotations were collected in our study, including the pilot runs. The average time to complete the task with the final template was 32 seconds.

B Human evaluation template

B.1 Golden set concept-attribute pairs

The concept - attribute pairs used for the golden set and the validation of the human evaluation template include: `<color, flower>`, `<material, container>`, `<color, language>`, `<background, animal>`, `<material, chair>`, `<side dish, cookie shape>`, `<pattern, clothing>`, `<style, building>`, `<weather, biome>`, `<color, vehicle>`.

B.2 User interface screenshots

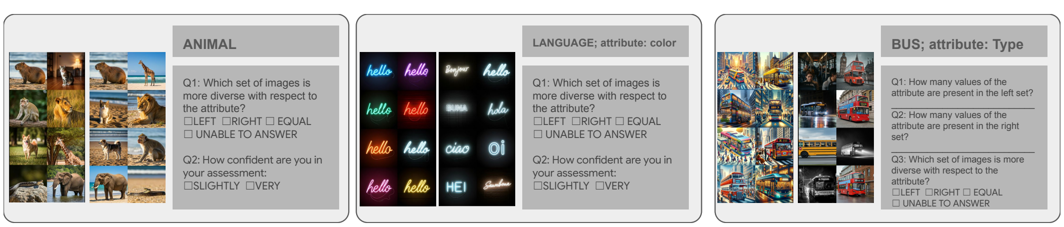


Figure 9: Examples of human evaluation templates used in the pilot study. In the template variant `w/o aspect`, only the `category` is provided. In the variant with `count`, an additional question is included for each set, prompting annotators to specify the number of distinct values observed for the target attribute within the corresponding image set. For exact examples see Figs. 10-12.

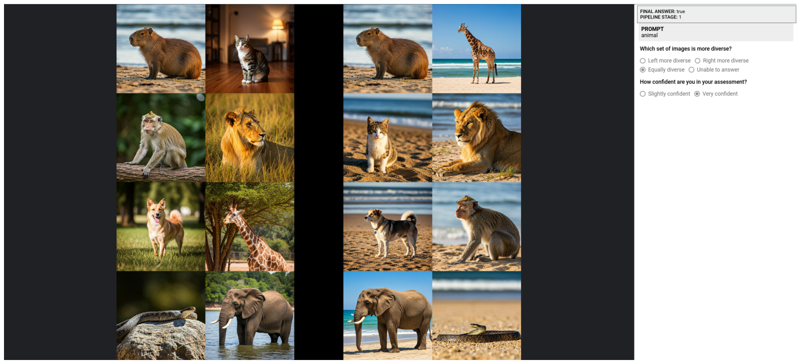


Figure 10: A screenshot of the user interface for one annotation example for the condition "No aspect".

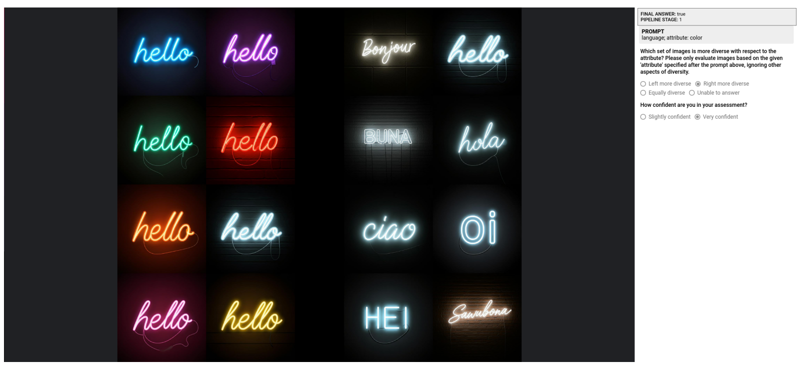


Figure 11: A screenshot of the user interface for one annotation example for the condition "Aspect".

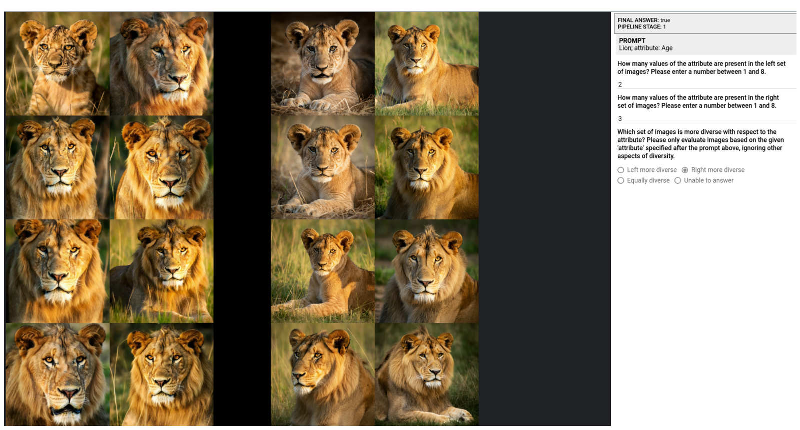


Figure 12: A screenshot of the user interface for one annotation example for the condition “Count”.

C Additional human evaluation results

In Fig. 13 we show the histogram of counts averaged across the 5 raters each set in all side-by-side comparisons.

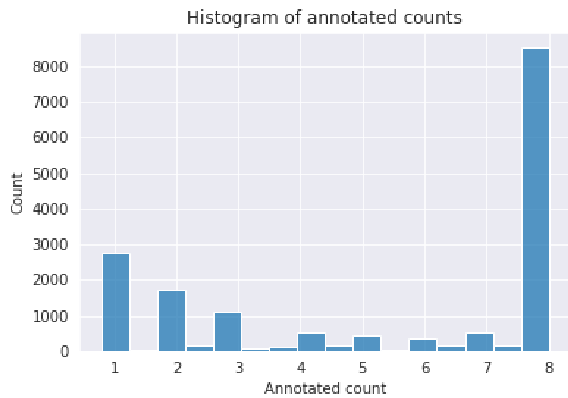


Figure 13: Distribution of all counts annotated by human raters.

D Additional autoevaluation results

D.1 Compute Usage

We used accelerators for running automatic evaluation metrics and generating the images. We run all metrics on a TPU V3 hardware[2]. The image generation pipeline ran on 4 TPUs.

D.2 Performance for detecting equally diverse image sets

We evaluate how good embeddings are at detecting equally diverse image sets. To not have a threshold-dependent metric, we use the area-under-the-ROC curve (AUC). We construct the true binary label as whether the image sets are labelled as equally diverse or not. We construct the scores as the absolute difference between the metric scores. We then plot the AUC. A good metric would have an AUC close to one, indicating that when the differences are small, the image sets are more likely to have been labelled as the same by the human annotators. We plot results in Figure 14, and find that no metric performs particularly well (AUC < 0.6 in all cases). However, the IMAGENET INCEPTION one performs best, presumably as it is trained to be invariant to small differences and so, as we can see in Figures 7-15, as a lack of diversity usually arises when images are very similar, the embedding performs well. However, we hypothesise that in the face of confounders (e.g. we want to measure diversity of the color of an object but not the type of object), we would not expect such an embedding to do well.

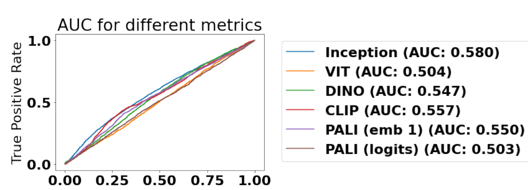


Figure 14: AUC to measure metrics ability to identify sets of equal diversity. It is clear that no metric is particularly effective at differentiating visually similar versus not sets of images.

[2] https://cloud.google.com/tpu/docs/v3

D.3 Additional qualitative results

In Fig. 15 we visualize examples for four side-by-side comparisons where the corresponding autoraters indicate that a group of images have highest or lowest diversity.

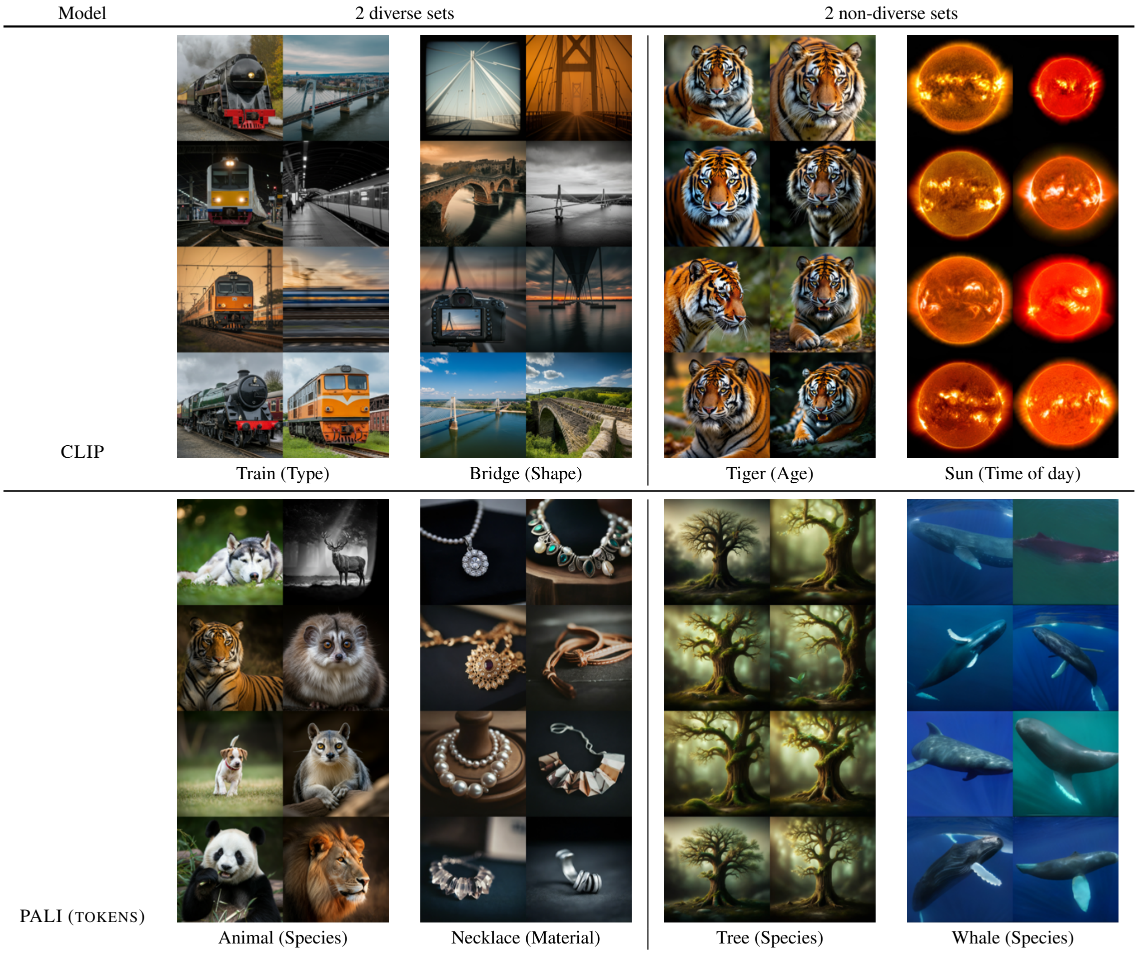


Figure 15: Qualitative results for different models, showing two very diverse and two non diverse sets.

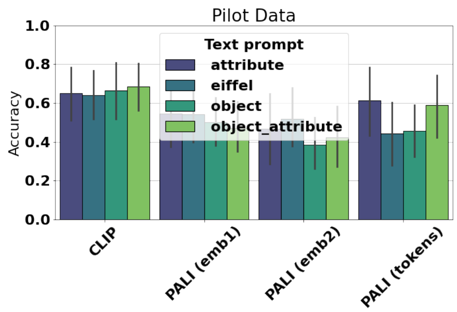
D.4 Impact of the prompt for the multimodal embeddings

We explore how the choice of prompt impacts results for the multimodal embeddings. We explore four different prompts which differ in their specificity and relatedness to the attributes under question. `[attribute]` and `[object]` are placeholders and filled in based on the object / attribute under test. The templates we consider are as follows:

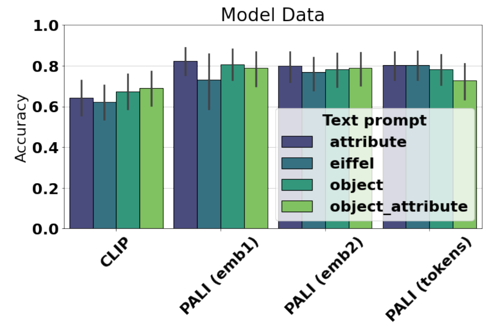
1. OBJECT_ATTRIBUTE: `What is the [attribute] of the [object]?`
2. ATTRIBUTE: `What is the [attribute]?`
3. OBJECT: `What is the [object]?`
4. EIFFEL: `Where is the Eiffel Tower?`

We would expect the first two questions to be most effective as they directly ask about the property for which we are measuring diversity. The object may be related but can be a confounder and the "Eiffel Tower" question is unrelated.

Results are shown in Figure 16. Surprisingly, we find that we do not see consistent benefit from the two most related prompts (OBJECT_ATTRIBUTE, ATTRIBUTE), implying that the embeddings are



(a) Results on the "diverse" golden set.



(b) Results on the annotation set, where annotators see count differences > 4.

Figure 16: Additional auto-eval results that show how results vary based on the textual prompt for the multimodal embeddings. We can see that we *do not* see consistently better results with more related prompts (`What is the [attribute] of the [object]?`, `What is the [attribute]?`), implying the textual input is being ignored.

mostly vision based. A more controllable multimodal embedding we hypothesise would be more effective in this setting.

D.5 Model ranking with autoevaluation approaches

In this section, we include more results for model ranking based on our auto-evaluation approaches:

- Figures 17, 18 and 19 show the results of compare model rankings in terms of significance in the number of wins with Wilcoxon signed-rank tests under a 95% confidence level using additional models to compute embeddings. This figure completes Figure 8 in Sec. 3.3. In theses figures, we can see:
  - Model ranking based on other embeddings. We observe that similarly to the observations in Sec. 3.3, for all embeddings except IMAGENET VIT, Imagen3 is not worse than all other models. We also observe that independently of the choice of embedding, Flux1.1, Imagen3 and DALLE3 are not worse than Muse2.2 and Imagen2.5. The differences between the models in the top group and the bottom group are more or less detected depending on the embeddings.
  - As mentioned in the main text, we also see the differences between multimodal models. These results highlight how the influence of the choice of embedding models and of conditioning on the model ranking results.
- Figures 20, 21 and 22 show the win rates corresponding to the results shown in Figure 8 in Sec. 3.3 and the additional results described above on the left panels, and compare the distributions of the two best and closest models in terms of behavior according to human evaluation, Imagen3 and Flux1.1, on the right panels. These figures correspond respectively to image models, multimodal model conditioned on attributes, and multimodal models conditioned on objects and attributes.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	>	=	>	>
Imagen 3	<	×	<	=	>
DALLE3	=	>	×	>	>
Muse 2.2	<	=	<	×	>
Imagen 2.5	<	<	<	<	×

(a) ViT embeddings.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	<	>	>	>
Imagen 3	>	×	>	>	>
DALLE3	<	<	×	>	>
Muse 2.2	<	<	<	×	>
Imagen 2.5	<	<	<	<	×

(b) DINO embeddings.

Figure 17: **Model ranking using auto evaluation approaches with additional image models.** We compare model rankings in terms of significance in the number of wins with Wilcoxon signed-rank tests under a 95% confidence level. Each entry in the each of the grids represents a comparison between two models. The $>$ sign indicates the model in the row is better, worse ($<$), or not significantly different (=) than the model in the column. The win rates in each of the grids are computed using the scores based on (a) IMAGENET VIT embeddings and (b) DINO embeddings.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	=	>	>	>
Imagen 3	=	×	>	=	>
DALLE3	<	<	×	=	=
Muse 2.2	<	=	=	×	=
Imagen 2.5	<	<	=	=	×

(a) CLIP embeddings.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	=	=	>	>
Imagen 3	=	×	=	>	>
DALLE3	=	=	×	>	>
Muse 2.2	<	<	<	×	>
Imagen 2.5	<	<	<	<	×

(b) PALI(emb2) embeddings.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	<	=	>	=
Imagen 3	>	×	>	>	>
DALLE3	=	<	×	>	>
Muse 2.2	<	<	<	×	=
Imagen 2.5	=	<	<	=	×

(c) PALI(tokens) embeddings.

Figure 18: **Model ranking using auto evaluation approaches with additional vision and language models conditioned on attributes.** We compare model rankings in terms of significance in the number of wins with Wilcoxon signed-rank tests under a 95% confidence level. Each entry in the each of the grids represents a comparison between two models. The $>$ sign indicates the model in the row is better, worse ($<$), or not significantly different (=) than the model in the column. The win rates in each of the grids are computed using the scores based on (a) CLIP embeddings, (b) PALI(emb2) embeddings, and (c) PALI(tokens) embeddings. All models are conditioned on attributes.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	<	>	>	>
Imagen 3	>	×	>	>	>
DALLE3	<	<	×	=	=
Muse 2.2	<	<	=	×	=
Imagen 2.5	<	<	=	=	×

(a) CLIP embeddings.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	<	=	>	>
Imagen 3	>	×	>	>	>
DALLE3	=	<	×	>	>
Muse 2.2	<	<	<	×	>
Imagen 2.5	<	<	<	<	×

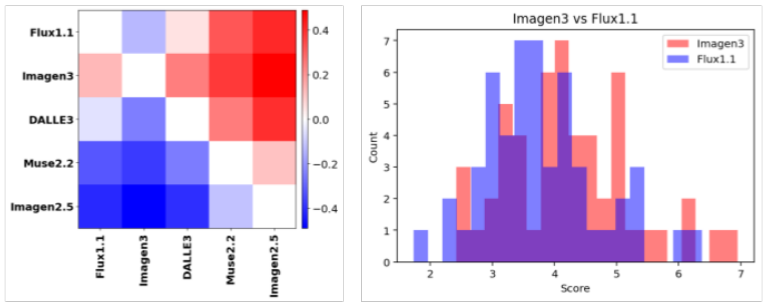
(b) PALI(emb2) embeddings.

	Flux 1.1	Imagen 3	DALLE3	Muse 2.2	Imagen 2.5
Flux 1.1	×	=	=	>	>
Imagen 3	=	×	=	=	>
DALLE3	=	=	×	=	>
Muse 2.2	<	=	=	×	=
Imagen 2.5	<	<	<	=	×

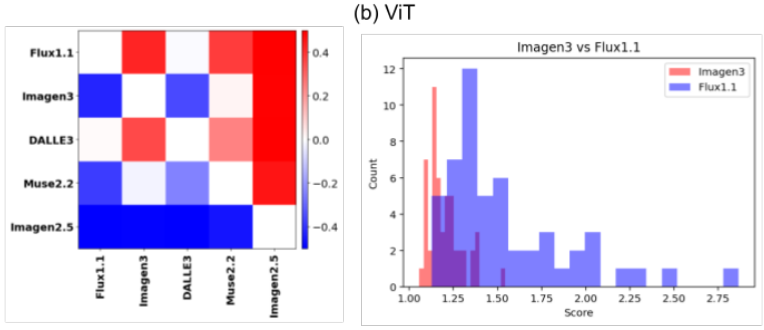
(c) PALI(tokens) embeddings.

Figure 19: **Model ranking using auto evaluation approaches with additional vision and language models conditioned on objects and attributes.** We compare model rankings in terms of significance in the number of wins with Wilcoxon signed-rank tests under a 95% confidence level. Each entry in the each of the grids represents a comparison between two models. The $>$ sign indicates the model in the row is better, worse ($<$), or not significantly different (=) than the model in the column. The win rates in each of the grids are computed using the scores based on (a) CLIP embeddings, (b) PALI(emb2) embeddings, and (c) PALI(tokens) embeddings. All models are conditioned on objects and attributes.

(a) Inception



(b) ViT



(c) DINO

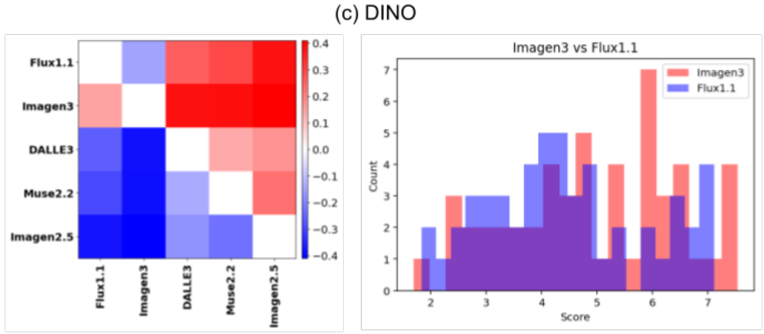


Figure 20: **Model ranking using auto evaluation approaches.** Win rate matrices and score distributions for Flux1.1 and Imagen3 using image models to compute embeddings.

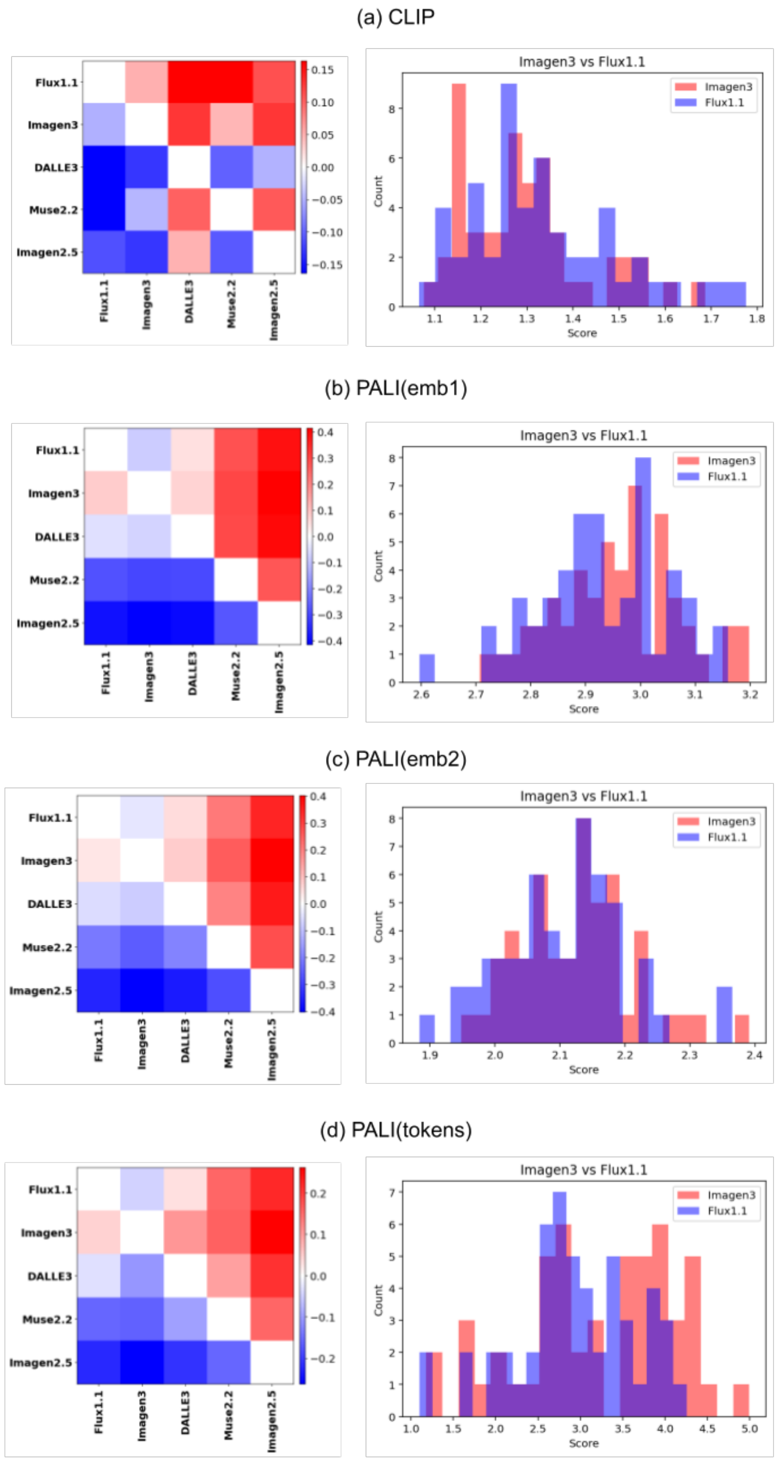


Figure 21: **Model ranking using auto evaluation approaches.** Win rate matrices and score distributions for Flux1.1 and Imagen3 using text-conditioned multimodal models to compute embeddings, conditioned on attributes.

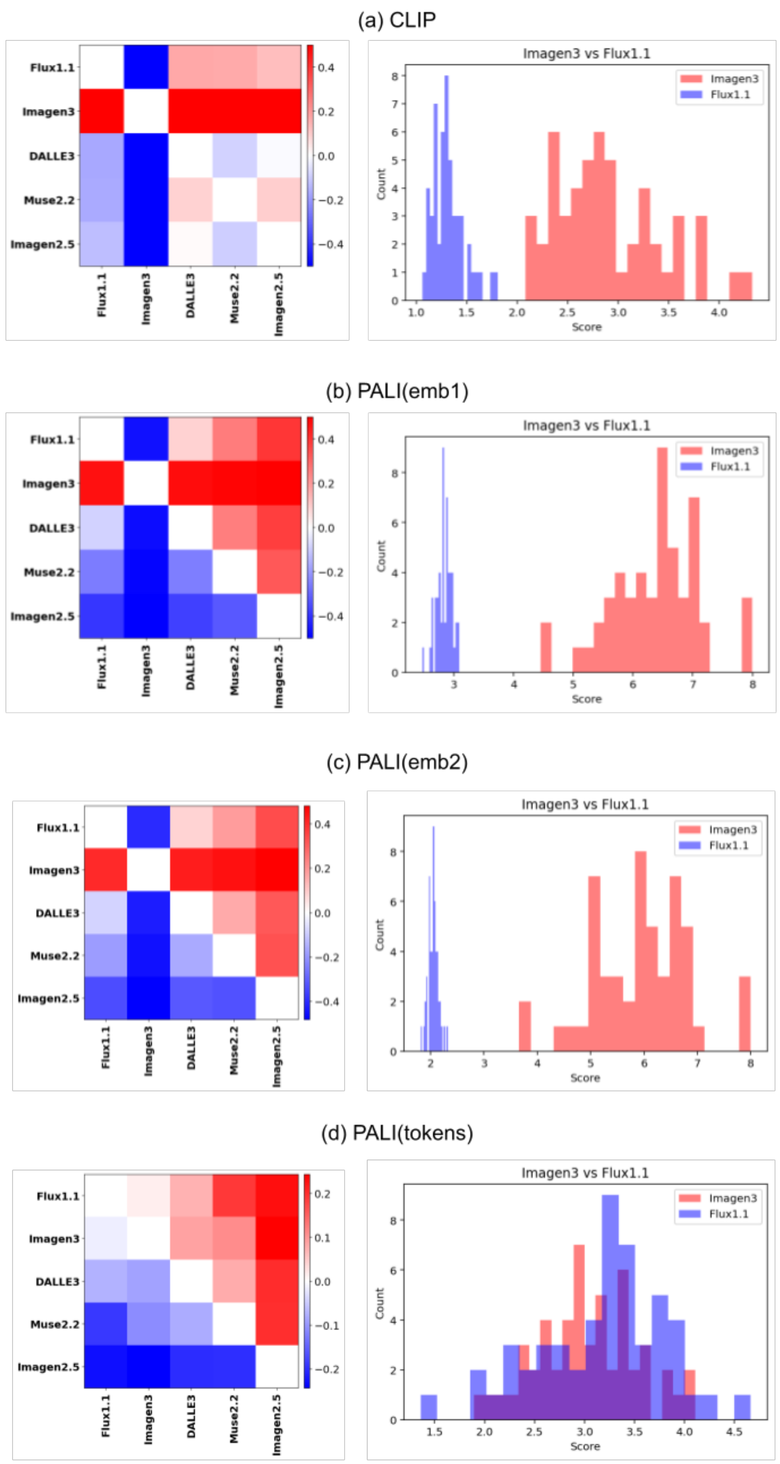


Figure 22: **Model ranking using auto evaluation approaches.** Win rate matrices and score distributions for Flux1.1 and Imagen3 using text-conditioned multimodal models to compute embeddings, conditioned on objects and attributes.

D.6 Evaluating diversity using foundation models

Besides the investigating multiple embeddings with the Vendi Score for evaluating diversity as presented in Sec. 3, we also propose to use the Gemini model family [37] for comparing T2I models in terms of attribute-based diversity. For that, we use the following instruction: "*I am currently comparing two models with the prompt [prompt] and I would like to know which model generates more diverse images with respect to the attribute [attribute], while disregarding any other attribute in the images. In the following image I show [number of images] images generated by one model in the left, which is [model in the left side] and [number of images] images generated by another model in the right, which is [model in the right side]. You must count the number of different instances of [attribute] in both sets and use this information to decide which set is the most diverse. If there is a set of images which is more diverse than the other with respect to [attribute], can you tell me which one is the most diverse set and explain why? Any other aspects in the images besides [attribute] must not be taken into account. You can also respond that both sets are equally diverse.*" In addition to the instruction, similarly to the human evaluation, two sets of images are given to the model as input.

In Fig. 23 we show the results of three different Gemini models on the task by showing the accuracy in the golden set described in Sec. 2.3. The most recent version of Gemini, v2.5 Flash, achieves the best performance, even surpassing the human raters in this task. These results indicate that such approaches are promising strategies for evaluating diversity which are: (i) able to capture cases where diversity is equally represented in both sets and (ii) do not rely on extracting embeddings.

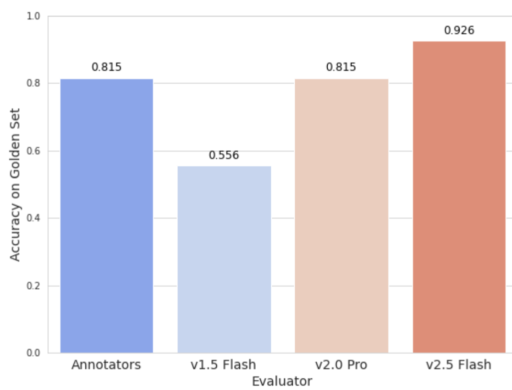


Figure 23: Accuracy of autoraters based on the Gemini model family on the task of comparing diversity of side-by-side sets of 8 images from the golden set. Most recent versions of Gemini perform better in the task, with the v2.5 Flash model surpassing the accuracy of human evaluators.