Do I look like a "cat.n.01" to you? A Taxonomy Image Generation Benchmark

Viktor Moskvoretskii[1,2], Alina Lobanova[3], Ekaterina Neminova[2,3], Alexander Panchenko[2,4], and Irina Nikishina[5]
[1]EPFL [2]Skoltech [3]HSE University [4]AIRI [5]University of Hamburg
vvmoskvoretskii@gmail.com

Abstract

This paper explores the feasibility of using text-to-image models in a zero-shot setup to generate images for taxonomy concepts. While text-based methods for taxonomy enrichment are well-established, the potential of the visual dimension remains unexplored. To address this, we propose a comprehensive benchmark for Taxonomy Image Generation that assesses models' abilities to understand taxonomy concepts and generate relevant, high-quality images. The benchmark includes common-sense and randomly sampled WordNet concepts, alongside the LLM generated predictions. The 12 models are evaluated using 9 novel taxonomy-related text-to-image metrics and human feedback. Moreover, we pioneer the use of pairwise evaluation with GPT-4 feedback for image generation. Experimental results show that the ranking of models differs significantly from standard T2I tasks. *Playground-v2* and *FLUX* consistently outperform across metrics and subsets and the retrieval-based approach performs poorly. These findings highlight the potential for automating the curation of structured data resources.

Figure 1: Generations of the *Playground* model from two kinds of input: a prompt taken from DiffusionDB (Wang et al., 2023), a dataset of prompts written for text-to-image models, and the inputs that WordNet-3.0 actually provides for a concept. The DiffusionDB prompt is far more detailed, and the model's internal representation of a WordNet concept can be misleading even when the definition is supplied.

1 Introduction

In recent years, Large Language Models (LLMs) and Visual Language Models (VLMs) have demonstrated remarkable quality across a wide range of single- and cross-domain tasks (Esfandiarpoor et al., 2024; Esfandiarpoor and Bach, 2023; Du et al., 2023; Jiang et al., 2024b). Their capabilities also expand to the tasks traditionally dominated by human input, such as annotation and data collection (Tan et al., 2024). At the same time, manually created datasets and databases remain in demand, as they are more accurate and reliable (Zhou et al., 2023), even though they are time-consuming and expensive to keep up to date.

In this paper, we focus on taxonomies — lexical databases that organize words into a hierarchical structure of "IS-A" relationships. WordNet (Miller, 1998) is the most popular taxonomy for English, forming the graph backbone for many downstream tasks (Mao et al., 2018; Lenz and Bergmann, 2023; Fedorova et al., 2024). In addition to textual data, taxonomies also extend to visual sources, e.g. ImageNet (Deng et al., 2009). ImageNet is built upon the WordNet taxonomy by associating concepts or "synsets" (sets of synonyms, aka lemmas) with thousands of manually curated images. However, it covers a very small portion of WordNet taxonomy (5,247 out of 80,000 synsets in total, 6.5%).

The coverage figure above is the scale of the problem: more than 74,000 WordNet synsets carry no image at all, and the models that could supply one have never been measured on the task. From

(a) Kandinsky 3 (b) SD3 (c) Playground (d) Retrieval

Figure 2: The example of a generation and retrieval results for *cigar lighter*. As can be observed, the generation approach is significantly superior to the retrieval approach, as the retrieved image is quite unconventional.

the visual perspective, Text-to-Image models are widely used for the visualizations (Ng et al., 2024; Sha et al., 2023), but only occasionally for taxonomies (Patel et al., 2024a). Therefore, there is limited knowledge about how well text-to-image models are capable of visualizing concepts of different level of abstraction in comparison to humans (Liao et al., 2024). Earlier work reaches the same conclusion from narrower angles: Baryshnikov and Ryabinin (2023) score generated images for hypernymy but depend on an ImageNet classifier, and Liao et al. (2024) cover abstract concepts alone. Image generation for taxonomies could be quite specific and require additional research: Figure 1 highlights the key differences in prompt usage for the DiffusionDB dataset (Wang et al., 2023) and WordNet-3.0. Moreover, the output taxonomy-linked depictions should aim to portray the synset's core idea succinctly, and sometimes to reveal insights about the concept that are hard to convey.

We address this gap by investigating the use of automated methods for updating taxonomies in the image dimension (depicting). Specifically, we develop a comprehensive evaluation benchmark comprising 9 metrics for Taxonomy Image generation using both human and automatic evaluation and a Bradley-Terry model ranking in line with recent top-rated evaluation methodology (Chiang et al., 2024a; Zheng et al., 2023b). Notably, our task produces model rankings that differ from those reported by established text-to-image benchmarks (Jiang et al., 2024a), demonstrating that it captures distinct and important capabilities. We also uncover that modern Text-to-Image models outperform traditional retrieval-based methods in covering a broader range of concepts, highlighting their ability to better represent and visualize these previously underexplored areas.

The contributions of the paper are as follows:

- We propose a benchmark comprising 9 metrics, including
  - several taxonomy-specific text-to-image metrics (grounded with theoretical justification drawing on KL Divergence and Mutual Information);
  - pairwise preference evaluation with GPT-4 for text-to-image generation and analyze its alignment with human preferences, biases, and overall performance.
- We test on the dataset specifically designed for Taxonomy Image Generation task, which presents challenges that were previously unaddressed in text-to-image research.
- We are the first to evaluate 11 open-weight Text-to-Image models and a retrieval baseline, 12 systems in total, on their ability to generate images for WordNet concepts.
- We publish the dataset of the images generated by the best Text-to-Image approach from the benchmark that fully covers WordNet-3.0 extending the ImageNet dataset.[1]

2 Related Work

In this section, we describe existing evaluation benchmarks for both texts and images and provide an overview of text-to-image generation models. We do not provide an overview on the existing taxonomy-related tasks and approaches and refer to Zeng et al. (2024) and Moskvoretskii et al. (2024b).

[1] https://github.com/s-nlp/TaxonomyImage

Please act as an impartial judge and evaluate the quality of the images provided by two AI image assistants to the user prompt displayed below.

You should choose the assistant that provides an image that follows the user's instructions better and reflects the user's prompt main concept better.
Your evaluation should consider factors such as image-text alignment, relevance, accuracy, depth, and fidelity (overall image quality).
Begin your evaluation by comparing the two images and provide a short explanation.

Avoid any position biases and ensure that the order in which the responses were presented do not influence your decision.

Do not allow the size of the images to influence your evaluation.
Do not favor certain names of the assistants.
Be as objective as possible.

After providing your explanation, output your final verdict by strictly following this format:

"[[A]]" if assistant A is better, "[[B]]" if assistant B is better, "[[C]]" for a tie and "[[D]]" if both images are bad.

[User Prompt]

inherited disorder (a genetic condition passed down from parents to their offspring, oftenresulting in physical or mental health problems))

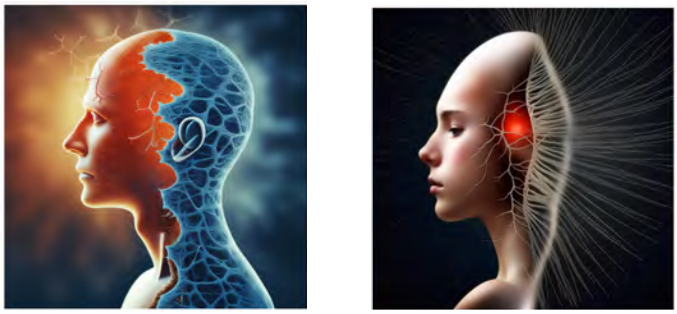


Figure 3: LLM prompt example for evaluating text-to-image assistants.

Evaluation Benchmarks Popular benchmarks for language models include GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a), MTEB (Muennighoff et al., 2023), SQuAD (Rajpurkar et al., 2016), MT-Bench (Zheng et al., 2023c) and others. For Text2Image Generation, there are benchmarks such as MS-COCO (Lin et al., 2014), Fashion-Gen (Rostamzadeh et al., 2018) or ConceptBed (Patel et al., 2024b). There are also platforms for interactive comparison of AI models, based on ELO-rating: LMSYS Chatbot Arena (Chiang et al., 2024b) for LLM and GenAI Arena (Jiang et al., 2024a) for comparing text-to-image models. Moreover, due to latest AI's abilities, "LLM-as-a-judge" evaluation emerged: text or image generation outputs of different models are compared by another model, see (Zheng et al., 2023c; Wei et al., 2024; Chen et al., 2024a).

Text-to-Image Generation Models For Taxonomies Image generation has recently received significant attention in the field of machine learning. Previously, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Autoencoders (VAEs) (Kingma and Welling, 2014) were primarily used for this purpose. However, diffusion-based methods have now become the dominant approach and are widely used for the visualizations (Ng et al., 2024; Sha et al., 2023), but rarely for taxonomies (Patel et al., 2024a).

To the best of our knowledge, the existing work on the evaluation of images for taxonomies comprises the paper of Baryshnikov and Ryabinin (2023), which introduces In-Subtree Probability (ISP) and Subtree Coverage Score (SCS), which are revisited in our paper. Recently, Liao et al. (2024) introduced a novel task of text-to-image generation for abstract concepts. The benchmark from Patel et al. (2024a) addresses grounded quantitative evaluations of text conditioned concept learners, and Zhang et al. (2024) also work with the notion of concepts for images when developing their concept forgetting and correction method. To the best of our knowledge, existing benchmarks do not explicitly evaluate concepts according to their position in the taxonomy. Our benchmark addresses this limitation by explicitly incorporating each concept's taxonomic position into the evaluation process.

3 Datasets

This section provides an overview of the datasets used to evaluate the performance of text-to-image (TTI) models. It includes the Easy Concepts dataset, the TaxoLLaMA test set derived from WordNet, and the predictions generated by the TaxoLLaMA model. The aim of the datasets is to assess the models' sensitivity to easier/harder dataset and to existing/AI-generated entities.

3.1 Easy Concept Dataset

The Easy Concepts dataset from Nikishina et al. (2023) comprises 22 synsets selected by the authors as common-sense concepts (e.g. *"coin.n.01, chromatic_color.n.01, makeup.n.01, furniture.n.01"*, etc.). We extend this list by including their direct hyponyms ("children nodes"), following the methodology outlined in the original paper and based on the English WordNet (Miller, 1998). The resulting dataset comprises 483 entities and represents a broader set of common knowledge entities.

3.2 Random Split from WordNet

To generate the second dataset, we use the algorithm from TaxoLLaMA (Moskvoretskii et al., 2024b). We randomly sample nodes connected to other synsets through the following types of hierarchical relationships in WordNet:

- **Hyponymy (Hypo)**: from a broader word (*"working_dog.n.01"*) to a more specific (*"husky.n.01"*). Here we take a broader word for image generation.
- **Hypernymy (Hyper)**: from a more specific word (*"cappuccino.n.01"*) to a broader concept (*"coffee.n.01"*). Here we take a more specific word for image generation.
- **Synset Mixing (Mix)**: nodes created by mixing at least two nodes (e.g. *"milk.n.01"* is a *"beverage.n.01"* and a *"dairy_product.n.01"*). Here we take the node created by mixing.

The algorithm for sampling uses a 0.1 probability for sampling Hyponymy, a 0.1 probability for sampling Synset Mixing, and a 0.8 probability for sampling Hypernymy. The dominance of Hypernymy is necessary because it is the most useful relation for training TaxoLLaMA (Moskvoretskii et al., 2024a). To mitigate this bias, the probabilities of occurrence in the test set differ between cases: for Hypernymy is set very low at 1×10^{-5}, higher for Hyponymy at 0.05, and highest for Synset Mixing at 0.1, as these cases are rare; these are per-node inclusion rates over pools of very different size to maintain real-world taxonomy structure.

The resulting test set includes 1,202 nodes: 828 from Hypernymy relations, 170 from Synset Mixing relations, and 204 from Hyponymy relations, mirroring their frequency in WordNet; every metric is also reported per case.

3.3 LLM Predictions Datasets

As our final goal is depicting of the new concepts for taxonomy extension, we should also test TTI models with LLM predictions rather than ground-truth synsets. Therefore, we finetune an LLM model on the Taxonomy Enrichment task to use its predictions and assess the sensitivity of text-to-image (TTI) models to AI-generated content.

The workflow comprises three steps: (i) exclude the Easy Concept dataset and random split from the overall WordNet data for LLM model training; (ii) train the updated version of TaxoLLaMA with LLaMA-instruct-3.1 (Grattafiori et al., 2024); (iii) solve the Taxonomy Enrichment task for the test data to generate concepts for visualization. When training the TaxoLLaMA-3.1 model, we follow the methodology in Moskvoretskii et al. (2024b), detailed in Appendix B.

This process resulted in 1,685 items. To match the original WordNet synsets, we generate definitions for every generated node with GPT-4.

4 Experimental Setup

In this section, we describe eleven TTI models and one Retrieval model (12 in total) and the details of image collection. Table 1 comprises the full list of the models compared in the evaluation benchmark, and each one is described in Appendix J.

An example of a prompt for image generation is demonstrated below, details are described in Appendix A. We perform experiments with two versions of the prompt: with and without definition. It is worth noting that adding definitions does not turn the task into "standard instruction following." In TTI settings, definitions are not a typical or natural form of instruction, therefore it is an additional informative diagnostic, revealing how models retrieve fine-grained taxonomic meaning. Figure 2 shows the example of a generation and retrieval results guided by the prompt from WordNet glosses:

`TEMPLATE: An image of <CONCEPT>`

Model name	Size	Model Family	Paper
SD-v1-5	400M	U-Net	Rombach et al. (2022)
SDXL	6.6B		Podell et al. (2024)
SDXL Turbo	3.5B		Liu et al. (2024)
Kandinsky 3	12B		Arkhipkin et al. (2023)
Playground-v2-aesthetic	2.6B		Li et al. (2023)
Openjourney	123M		Prompthero (2023)
IF	4.3B	Diffusion Transformers	DeepFloyd.Lab (2023)
SD3	2B		Esser et al. (2024)
PixArt-Sigma	900M		Chen et al. (2024b)
Hunyuan-DiT	1.5B		Li et al. (2024)
FLUX	12B		BlackForestLabs (2024)
Wikimedia Commons[2]	-	Retrieval	Ferrada et al. (2018) Jones and Oyen (2022)

Table 1: List of the approaches evaluated in the Taxonomy Image Generation benchmark.

```
(<DEFINITION>)
  EXAMPLE: An image of cigar lighter (a
lighter for cigars or cigarettes)
```

5 Evaluation

Our evaluation consists of 9 metrics that we assess to provide a comprehensive evaluation using the latest methods. To formally define our metrics, let V be a finite set of concepts v, $A(v) \subseteq V$ the set of hypernyms for v, and $N(v) \subseteq V$ the set of cohyponyms for v. Let X^j be the set of all possible images x^j in a finite model space $j \in J$ and $|J| = 12$ in our case. We define a mapping $g^j : V \rightarrow X^j$ for each model j, which assigns an image to each concept v.

5.1 Preferences Metrics

ELO Scores We evaluate the ELO scores of the model by first assigning pairwise preferences, similar to the modern evaluation of text models (Chiang et al., 2024a). Each object v is assigned two uniformly sampled random models $A, B \sim \mathbb{U}[J]$, and their outputs (x^A, x^B) engage in a battle. Then, either a human assessor or GPT-4 serves as a function to assign a win to model A (0) or model B (1), represented as $f(v, x^A, x^B) \in \{0, 1\}$. Ties are omitted in both the notation and the BT model. To compute ELO scores, the likelihood is maximized with respect to the BT coefficients for each model, which correspond to their ELO scores. The full procedure, from pairing to bootstrap, is given in Appendix D, following the Chatbot Arena methodology (Chiang et al., 2024a).

Following this methodology, we calculate the ELO score based on the Bradley-Terry (BT) model with bootstrapping to build 95% confidence intervals. The Bradley-Terry model (Bradley and Terry, 1952) is a probabilistic framework used to predict the outcome of pairwise comparisons between items or entities. It assigns a latent strength parameter π to each item i and the probability that item i is preferred over item j is given by: $P(i > j) = \frac{\pi_i}{\pi_i + \pi_j}$. Here, $\pi_i, \pi_j > 0$ represent the strengths of items i and j, respectively. The parameters are typically estimated from observed comparison data using maximum likelihood estimation. We also adopt a labeling technique that includes the "Tie" and "Both Bad" categories, indicating cases where the models are equally good or both produce poor outputs. We modify the prompt from previous studies evaluating text assistants (Zheng et al., 2023a) for images, as presented in Figure 3 (also see prompt 8 in Appendix E for more details).

GPT-4 is only one of the nine metrics we report, and it is used as a single comparative signal, not as the core evaluation mechanism. Therefore, we also conduct the **Human ELO Evaluation** along with the **GPT-4 ELO Evaluation** on 3370 pair images from two different models ($\approx$ 600 samples from each model). For Human ELO, we employ 4 assessors expert in computational linguistics, both male and female with at least bachelor degrees. The Spearman correlation between annotators is 0.8 (p-value ≤ 0.05) for the images generated with definitions. For the automatic calculation of the ELO score we use GPT-4, which is highly correlated with human evaluations (Zheng et al., 2023a) and has proven to be an effective image evaluator on its

own (Cui et al., 2024), as well as a great pairwise preferences evaluator (Chen et al., 2024a).

Reward Model We utilize the ImageReward model from a recent study (Xu et al., 2023), which is trained to align with human feedback preferences, focusing on text-image alignment and image fidelity. This score demonstrates a strong correlation with human annotations and outperformed the CLIP Score and BLIP Score. Formally, this metric is similar to ELO Scores, as the reward model was tuned using preferences and the BT model. However, the key difference is that each object v is assigned a real-valued score and takes only one model image x^j as input: $f_{reward}(v, x^j) \in \mathbb{R}$.

5.2 Similarities

In this section, we introduce novel similarity metrics that leverage taxonomy structure and are derived from KL Divergence and Mutual Information. They all have CLIP similarities under the hood, which have been already validated against human judgements (Hessel et al., 2021). This ensures that our metrics, by extension, are aligned with human judgements.

In practice, we approximate the probabilities using CLIP similarity (Hessel et al., 2021), as it is the most reliable measure of text-image co-occurrence.

Formally, CLIP model $C(\text{text or image}) \in \mathbb{R}^{hidden\_dim}$, we calculate the cosine similarity between the embedding of concept v and the embedding of image x^j, resulting in the score $\text{sim}(C(v), C(x^j))$

Lemma Similarity reflects how well the image aligns with the lemma's textual description, defined as

$$\begin{aligned} S_{\text{lemma}}(v, x) &:= P(X = x \mid v) \\ &\approx \text{sim}(C(v), C(x^j)). \end{aligned} \tag{1}$$

Hypernym Similarity reflects how similar the image is on average to the lemma hypernyms; is defined as

$$\begin{aligned} S_{\text{hyper}}(v, x) &:= P(X = x \mid A(v)) = \\ &\frac{1}{|A(v)|} \sum_{a \in A(v)} P(X = x \mid a) \approx \\ &\frac{1}{|A(v)|} \sum_{a \in A(v)} \text{sim}(C(a), C(x)). \end{aligned} \tag{2}$$

Cohyponym Similarity measures how similar the image is, on average, to the cohyponyms; is defined as

$$\begin{aligned} S_{\text{cohyponym}}(v, x) &:= P(X = x \mid N(v)) = \\ &\frac{1}{|N(v)|} \sum_{n \in N(v)} P(X = x \mid n) \approx \\ &\frac{1}{|N(v)|} \sum_{n \in N(v)} \text{sim}(C(n), C(x)). \end{aligned} \tag{3}$$

This metric should be interpreted in conjunction with Specificity, as a high Cohyponym Score paired with low Specificity does not necessarily indicate good generation.

In the T2I domain, it is not feasible to define "accuracy" in the traditional sense. It is difficult to determine whether the reflection of a concept is entirely correct or completely incorrect. This challenge is inherent to the nature of T2I tasks and is shared by other studies in this domain. To address this limitation, we propose an analogous measure to assess how well the image reflects the concept. We use the probability of the concept with respect to the generated image, denoted as $P(X = x|v)$, which is derived from Lemma Similarity. Hypernym and Cohyponym Similarity extend it to the node's WordNet neighbours: the image is scored against the *text* of those neighbours, so the context is lexical and every comparison is text-to-image. All three probabilities are approximated by the CLIP cosine similarity above; Appendix K gives the formal definitions.

In order to understand how well hypernym and co-hyponym similarities correlate with human semantic understanding, we calculated Spearman correlation of the model ranks assigned by these metrics and ranks assigned through the human evaluation ($\rho \approx 0.911, p \leq 0.00004$ for Hypernym CLIP-Score, $\rho \approx 0.871, p \leq 0.00022$ for Co-hyponym CLIP-Score). This demonstrates that the proposed metrics capture relations that humans reliably recognize. WordNet is manually curated, so both the concept texts and the hypernym and cohyponym links we score against are human-validated.

Specificity helps to ensure that the image accurately represents the lemma rather than its cohyponyms with the relation of the CLIP-Score to the Cohyponym CLIP-Score $\frac{S_{\text{hyper}}(v,x)}{S_{\text{cohyponym}}(v,x)}$

This metric generalizes the In-Subtree Probability, as proposed in (Baryshnikov and Ryabinin, 2023). The key advantage of our metric is that it does not depend on a specific ImageNet classifier and can be applied to any type of taxonomy node.

Metric	Mean	Ground Truth				Predicted			
		Easy	Hypo	Hyper	Mix	P-Easy	P-Hypo	P-Hyper	P-Mix
ELO GPT (w/ def)	Playground	Playground	Playground	Playground	Playground	PixArt	PixArt	Playground	SD3*
ELO GPT (w/o def)	Playground	Kandinsky3	PixArt*	Playground	FLUX	FLUX	Playground	Playground	Kandinsky
ELO Human (w def)	FLUX	Playground / DeepFloyd	Kandinsky3	FLUX	Playground	FLUX	FLUX	Playground	PixArt
ELO Human (w/o def)	FLUX	SD3	Kandinsky3	Playground	PixArt	FLUX	FLUX	SDXL	SDXL
Reward Model (w/ def)	Playground	Playground	Playground	Playground	Playground	Playground	Playground	Playground	Playground
Reward Model (w/o def)	Playground	Playground	Playground	Playground	Playground	Playground	Playground	Playground	Playground
Lemma Similarity	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo
Hypernym Similarity	SDXL-turbo	FLUX	FLUX / SDXL-turbo	SDXL-turbo	FLUX / SDXL-turbo	SDXL-turbo	SDXL-turbo	Draw	Draw
Cohyponyms Similarity	SDXL-turbo	FLUX	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo	SDXL-turbo / SDXL
Specificity	SD1.5	SD1.5 / Playground	SD1.5	SD1.5 / Playground	Draw	Draw	SDXL-turbo / SDXL	SDXL-turbo	SDXL-turbo
FID	SD1.5	FLUX	FLUX	FLUX	HDiT	FLUX	FLUX	FLUX	DeepFloyd
IS	SD3	PixArt	Playground	Playground	SD3	SD3	FLUX	FLUX / Retrieval	Playground

Table 2: Summary of the Top-1 model for each metric and subset. Each cell shows the best-rated model. If two models tie, both are listed with a slash; if more than two tie, "Draw" is written, indicating insufficient specificity. Results marked with * have negligible differences within the confidence interval. Subsets and models are described in Sections 3 and 4.

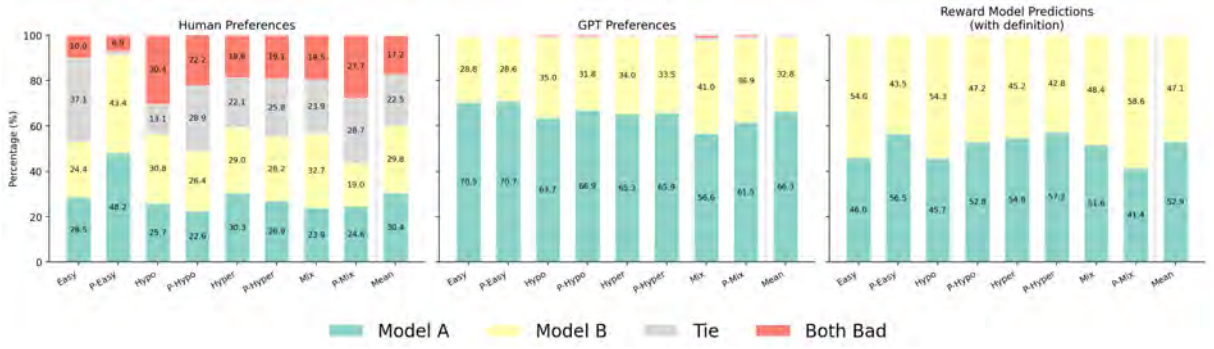


Figure 4: Distribution of preferences for Human and GPT across subsets in percentage (with definition).

5.3 FID and IS

We evaluate the Inception Score (IS) (Salimans et al., 2016) and the Fréchet Inception Distance (FID) (Heusel et al., 2017). IS is primarily used to assess diversity, while FID measures image quality relative to true image distributions. In our case, we calculate FID based on retrieved images, meaning that in this specific setting, FID reflects the "realness" or closeness to retrieval rather than the semantic correctness of an image.

6 Results

The summary of the main results is presented in Table 2 : they show the best model for each subset and each metric. Additionally, we provide an error analysis in Appendix F and discuss the strengths and weaknesses of the best-performing models. Concepts with no direct visual referent are still depicted, through rendered text, abstract imagery or a symbolic scene, and Appendix F analyses those cases.

ELO Scores are shown in Figure 5. FLUX and Playground take the top two places under both human and GPT-4 assessors and PixArt the third, while the middle of the ranking is less stable; the rank correlation between the two judges is 0.92, p-value ≤ 0.05. Without definitions (Figure 19, Appendix H) FLUX leads for humans and Playground for GPT-4, though overlapping intervals leave PixArt a contender, and the correlation falls to 0.73 while staying strong.

Agreement disappears at the level of individual battles, where GPT-4 shows a strong bias toward the first option that humans do not (Figure 4, and the confusion matrix in Figure 22, Appendix H). Most models benefit from definitions in the prompt.

Reward Model The reward model of Xu et al. (2023) prefers Playground, then PixArt and FLUX with no significant gap between the two, and prefers Playground on every subset as well (Figures 21 and 23, Appendix H). It correlates strongly with human evaluations (0.79) and only moderately with GPT-4 (0.59).

Similarities for lemmas, hypernyms and cohyponyms put SDXL-turbo ahead across subsets, and FLUX on Easy Ground Truth, with a stable ordering across the three metrics (Appendix I). This diverges from the preference metrics, most likely because CLIP-Score sees text-image alignment and not image quality. SDXL-turbo also outranks SDXL despite being its distilled version.

Specificity has no clear winner and varies little across models (Appendix I), with SDXL-turbo, SD1.5, and Playground leading. Although SD1.5 ranks low on preference, it leads several subsets, while Playground performs strongly on both pref-

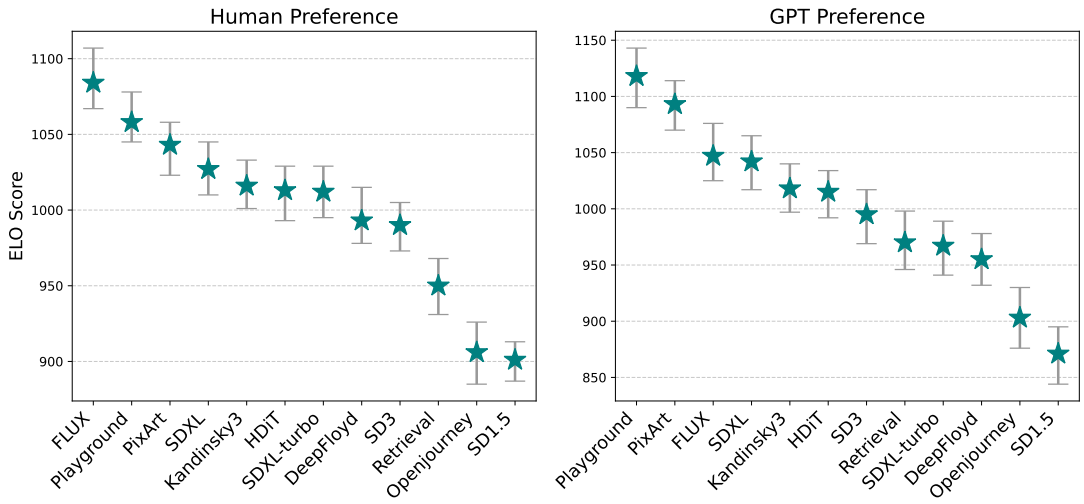


Figure 5: ELO scores for human and GPT-4 preferences. The prompt includes the definition. Overall Spearman correlation of model rankings remains significantly high at 0.92, p-value ≤ 0.05.

erence and lemma-level specificity.

FID favors SD1.5 on average, while FLUX leads on nearly every subset (Tables 4 and 6, Appendix I). This suits a model that stays close to crawled images and still appeals to human judgment.

IS puts SD3, Playground and Retrieval first across subsets, so their generations read as sharper and more distinct (Tables 4 and 5, Appendix I). However, incorporating definitions does not improve the performance of either the SDXL or SD3 model family.

Overall results show that Playground and FLUX are among the top models across different metrics, both with and without definitions. While PixArt also demonstrates strong results, it is preferred by AI evaluations more than human preferences, indicating that the preference may be more AI-Judge specific. Specificity is the least consistent axis: the SD family moves around across metrics and subsets, so CLIP-aligned training does not guarantee that an image pins the precise concept rather than its neighbours.

7 Fine-tuned Reward Model

To reduce reliance on costly, variable, and position-sensitive GPT-4 evaluations, we fine-tune a CLIP ViT-L/14 backbone on our human assessments with a Bradley-Terry pairwise loss, using only the decisive *A wins* and *B wins* labels, since *Tie* and *Both Bad* are underrepresented to support a four-way signal. Splits are separated at the synset level so no WordNet identifier is shared between them, and every pair appears in both presentation orders. Experiments were conducted both with and without definitions. Full architecture, hyperparameters, split sizes details, abd the results without definitions are provided in Appendix C. We use the off-the-shelf ImageReward model (Xu et al., 2023) evaluated in Section 5 as a general-purpose baseline. Comparing against this model allows us to assess whether fine-tuning on taxonomy-specific human preferences improves reward-based evaluation.

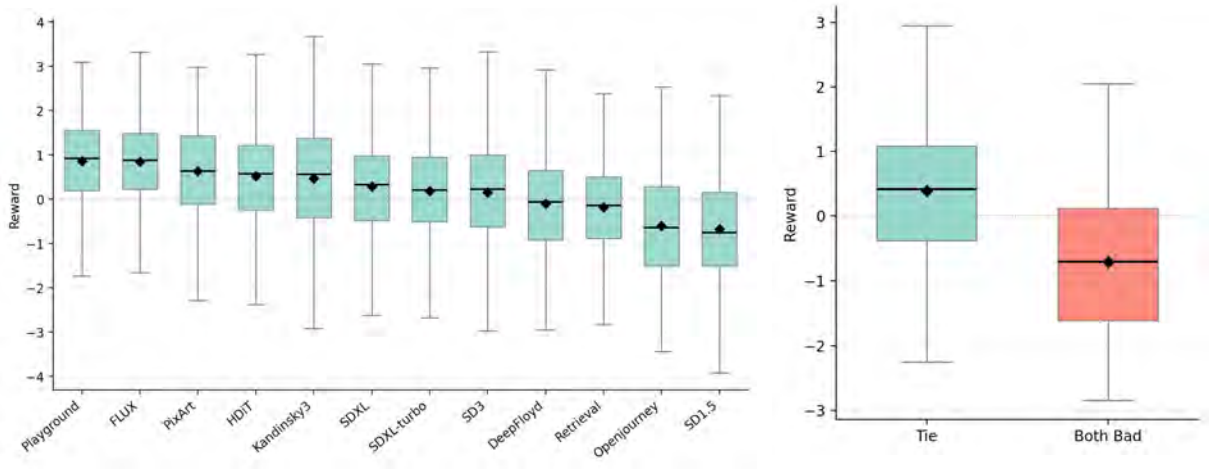
Results. Table 3 reports pairwise accuracy and F1 score on the test set of *A wins* / *B wins* pairs. The fine-tuned CLIP substantially outperforms the off-the-shelf ImageReward (Xu et al., 2023) in both the with- and without-definition settings.

Model	w/def	Accuracy	F1
ImageReward	yes	0.624	0.619
ImageReward	no	0.597	0.597
Fine-tuned CLIP	yes	**0.712**	**0.706**
Fine-tuned CLIP	no	0.698	0.697

Table 3: Comparison between baseline reward model and fine-tuned one.

Across the 12 evaluated generators, the ranking based on mean reward strongly agrees with the human ELO ranking, yielding a Spearman correlation $\rho = 0.94$ with human ELO scores (95% CI: $[+0.916, +0.958]$, $p \leq 0.05$). The three lowest-ranked models appear in exactly the same order, while the remaining discrepancies are limited to local swaps within narrow ELO clusters.

The suggested ordering (for the models with def-



(a) The distribution of rewards across TTI models.

(b) Distribution of Tie and Both Bad rewards. Mann-Whitney U, $p < 0.001$

Figure 6: The distribution of rewards generated by fine-tuned CLIP model. Prompt includes definition.

initions) is visualised in Figure 6a. Median rewards decrease monotonically across the ranking, and the two weakest models are clearly separated from the others by strongly negative scores. The first place is shared by Playground and FLUX, which is consistent with human annotations.

Figure 4 compares empirical preference distributions across subsets. GPT-4 exhibits a noticeable first-position bias, whereas our fine-tuned reward model distributes decisive votes almost evenly between *Model A* and *Model B*. This behaviour closely matches the symmetric distribution observed in the human annotations.

Although *Tie* and *Both Bad* pairs were excluded from training, the model still demonstrates coherent behaviour on these categories. Figure 6b shows that the reward distribution for *Tie* pairs is centered slightly above zero, while *Both Bad* pairs receive lower rewards and exhibit a broader spread. Although the distributions partially overlap, *Both Bad* pairs consistently obtain lower scores overall. This suggests that the reward model captures meaningful distinctions between neutral-quality and uniformly poor generations even without explicit supervision for these categories.

The model trained without definitions exhibits the same overall trend; further details are provided in Appendix C.

8 Conclusion

We have proposed the Taxonomy Image Generation benchmark as a tool for the further evaluation of text-to-image models in taxonomies, as well as for generating images in existing and potentially automatically enriched taxonomies. It consists of 9 metrics and evaluates the ability of 12 open-source text-to-image models to generate images for taxonomy concepts. Our evaluation results show that Playground (Li et al., 2023) ranks first in all preference-based evaluations.

Limitations

- Our evaluation focuses on open-source text-to-image models, as they are more convenient and cost-effective to use in any system than models relying on an API. Additionally, open-source models offer the flexibility for fine-tuning, which is not possible with closed-source models. However, it would be valuable to explore how closed-source models perform on this task, as our benchmark depends solely on the quality of the generated images.

- Preferences using GPT-4 were obtained with the use of Chain of Thought reasoning, following previous studies to optimize the prompt (Zheng et al., 2023a). However, we did not utilize multiple generations with a majority vote to improve consistency, nor did we rename the models, which could help reduce positional bias. Additionally, we did not perform multiple runs with models alternating positions, as each model could appear in position A or B with equal probability. The Bradley-Terry model of preferences compensates for such inconsistencies and provides robust scoring, given a sufficient number of preference labels, as noted in a previous study (Zheng et al.,

2023a). Our assumption is further supported by the high correlation in the resulting rankings, even though the correlation between raw preferences is close to zero.

- Metrics based on CLIP-Score may be biased toward the CLIP model and lack specificity if CLIP is unfamiliar with or unspecific to the precise WordNet concept. Additionally, models could be fine-tuned to optimize for this particular metric. To mitigate this bias, we propose incorporating preferences from AI feedback and also employ preferences from human feedback to provide a more balanced and comprehensive evaluation.

- The Inception Score relies on the InceptionV3 model, which is specific to ImageNet1k. We included this metric as it is traditionally used to measure overall text-to-image performance. However, to address this potential bias, we introduced CLIP based metrics as well as generalization of the ISP metric from a previous study (Baryshnikov and Ryabinin, 2023), which also relies on an ImageNet1k classifier, but we supplemented it with the use of CLIP to provide a broader evaluation.

- The benchmark focuses mainly on WordNet concepts, which may limit generalization to other (multilingual) taxonomies or domains that differ in structure. Extending the benchmark to other resources (e.g. Wikidata, ConceptNet) would require substantial additional design and alignment work and is therefore a separate research direction. In the present paper, we intentionally focus on WordNet because it provides the clearest foundation for linguistic evaluation of model representations.

Ethical Considerations

In our benchmark, we utilized several text-to-image models as well as text assistants for evaluating and generating new concepts. While these models are highly effective for creating creative and novel content, both textual and visual, they may exhibit various forms of bias. The models tested in this benchmark have the potential to generate malicious or offensive content. However, we are not the creators of these models and focus solely on evaluating their capabilities; therefore, the responsibility for any unfair or malicious usage lies with the users and the models' authors.

Additionally, our fine-tuning of the LLaMA 3.1 model was conducted using safe prompts sourced from WordNet. Although applying quantization and further fine-tuning could potentially reduce the model's safety, we did not observe any unsafe or offensive behavior during our testing.

Reproducibility

We publish all datasets, generated WordNet images and collected preferences in the GitHub repository.[3]

References

Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. 2023. Kandinsky 3.0 technical report. *arXiv preprint arXiv:2312.03511*.

Anton Baryshnikov and Max Ryabinin. 2023. Hypernymy understanding evaluation of text-to-image models via WordNet hierarchy. *arXiv preprint arXiv:2310.09247*.

BlackForestLabs. 2024. Announcements.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024b. PIXART-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXII*, volume 15090 of *Lecture Notes in Computer Science*, pages 74–91. Springer.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024a. Chatbot arena: An open platform for evaluating LLMs by human preference. *Preprint*, arXiv:2403.04132.

[3] https://github.com/VityaVitalich/TaxoGen

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024b. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xiao Cui, Qi Sun, Wengang Zhou, and Houqiang Li. 2024. Exploring GPT-4 vision for text-to-image synthesis evaluation. In *The Second Tiny Papers Track at ICLR 2024*.

DeepFloyd.Lab. 2023. DeepFloyd IF.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. 2023. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 9156–9172. Curran Associates, Inc.

Reza Esfandiarpoor and Stephen H Bach. 2023. Follow-up differential descriptions: Language models resolve ambiguities for image classification. *arXiv preprint arXiv:2311.07593*.

Reza Esfandiarpoor, Cristina Menghini, and Stephen Bach. 2024. If CLIP could talk: Understanding vision-language model representations through their preferred concept descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9797–9819, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.

Sebastián Ferrada, Nicolás Bravo, Benjamin Bustos, and Aidan Hogan. 2018. Querying wikimedia images using wikidata facts. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1815–1821, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *Advances in neural information processing systems*, 30.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. 2024a. GenAI arena: An open evaluation platform for generative models. *Preprint*, arXiv:2406.04485.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024b. A survey on large language models for code generation. *Preprint*, arXiv:2406.00515.

Shawn M Jones and Diane Oyen. 2022. Abstract images have different levels of retrievability per reverse image search engine. In *European Conference on Computer Vision*, pages 203–222. Springer.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Mirko Lenz and Ralph Bergmann. 2023. Case-based adaptation of argument graphs with wordnet and

large language models. In *Case-Based Reasoning Research and Development*, pages 263–278, Cham. Springer Nature Switzerland.

Daiqing Li, Aleks Kamko, Ali Sabet, Ehsan Akhgari, Linmiao Xu, and Suhail Doshi. 2023. Playground v2.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, and 26 others. 2024. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *Preprint*, arXiv:2405.08748.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2024. Text-to-image generation for abstract concepts. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 3360–3368. AAAI Press.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Bochao Liu, Pengju Wang, and Shiming Ge. 2024. Learning differentially private diffusion models via stochastic adversarial distillation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VII*, volume 15065 of *Lecture Notes in Computer Science*, pages 55–71. Springer.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024a. TaxoLLaMA: WordNet-based model for solving multiple lexical sematic tasks. *arXiv preprint arXiv:2403.09207*.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024b. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. 2024. PartCraft: Crafting creative objects by parts. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX*, volume 15067 of *Lecture Notes in Computer Science*, pages 420–437. Springer.

Irina Nikishina, Polina Chernomorchenko, Anastasiia Demidova, Alexander Panchenko, and Chris Biemann. 2023. Predicting terms in IS-a relations with pre-trained transformers. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 134–148, Nusa Dua, Bali. Association for Computational Linguistics.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024a. ConceptBed: Evaluating concept learning abilities of text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14554–14562.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024b. ConceptBed: Evaluating concept learning abilities of text-to-image diffusion models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 14554–14562. AAAI Press.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Prompthero. 2023. Openjourney.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *Preprint*, arXiv:1806.08317.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2023. DE-FAKE: detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, pages 3418–3432. ACM.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *Preprint*, arXiv:2402.13446.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911, Toronto, Canada. Association for Computational Linguistics.

Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. 2024. Systematic evaluation of LLM-as-a-Judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. *Preprint*, arXiv:2408.13006.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 15903–15935. Curran Associates, Inc.

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Zhenyu Wu, Shangbin Feng, and Meng Jiang. 2024. CodeTaxo: Enhancing taxonomy expansion with limited examples via code language prompts. *Preprint*, arXiv:2408.09070.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2024. Forget-me-not: Learning to forget in text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 1755–1764. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023a. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023c. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Technical Details

For text-to-image, we used the recommended generation parameters for each model, the HuggingFace Diffusers library, and a single NVIDIA A100 GPU. All models were utilized in FP16 precision and produced images with resolutions of 512x512 or 1024x1024. Additionally, we experimented with prompting, adding definitions from the WordNet database to help with ambiguity resolution, as this has shown benefits for LLMs in the past (Moskvoretskii et al., 2024b).

B TaxoLLaMA-3.1 Training

We follow the recipe of Moskvoretskii et al. (2024b) and change only the base model, so this appendix restates that recipe rather than sending the reader to the original paper.

Data. Hyponym-hypernym edges are sampled at random from the WordNet-3.0 graph, over nouns and verbs. A child node linked to several hypernyms contributes one pair per link. The child node's WordNet definition is included to disambiguate its sense. For our benchmark we first remove the Easy Concept dataset and the random split from the WordNet data, so that no node under evaluation is seen during training.

Prompt format. Each example is a single instruction turn, with the loss computed on the target tokens only and not on the instruction:

```
[INST] <<SYS>> You are a helpful assistant.
List all the possible words divided with
a comma. Your answer should not include
anything except the words divided by a
comma <</SYS>>
hyponym: tiger (large feline of forests in
most of Asia having a tawny coat with black
stripes) | hypernyms: [/INST]
big cat
```

Optimisation. The model is quantised to 4 bits and fine-tuned with QLoRA (Dettmers et al., 2023), a full-precision LoRA adapter (Hu et al., 2021). Training uses a batch size of 32, a learning rate of 3×10^{-4}, and a cosine annealing schedule. Moskvoretskii et al. (2024b) report the procedure to be markedly sensitive to the learning rate and the scheduler, and their runs fit on a single A100.

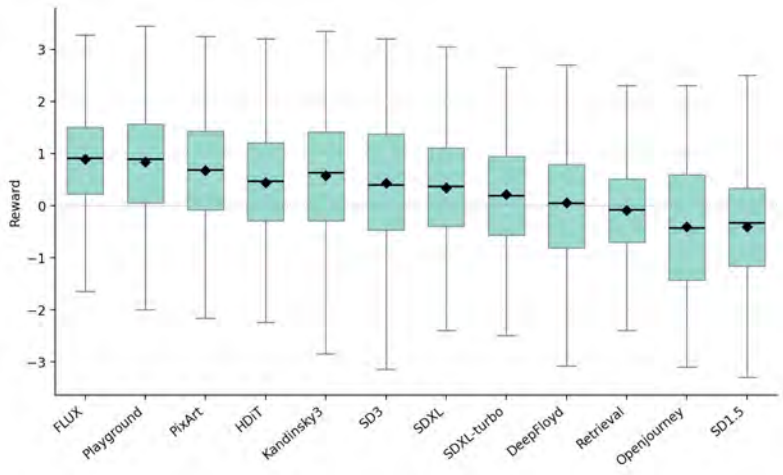
Difference from the original. Moskvoretskii et al. (2024b) build on LLaMA-2-7b; we substitute LLaMA-3.1-instruct (Grattafiori et al., 2024). Applying the resulting model to the Taxonomy Enrichment task yields the 1,685 predicted concepts described in Section 3.

C Fine-tuned Reward Model Details

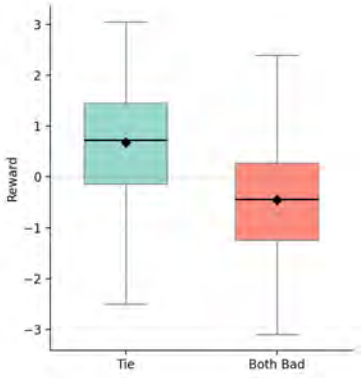
This appendix holds the setup for the reward model of Section 7, moved out of the main body for space.

Data. Training is restricted to the decisive *A wins* and *B wins* labels, as the *Tie* and *Both Bad* categories are substantially under-represented relative to the decisive classes, and the resulting class imbalance prevents reliable learning of a four-way preference signal at the current dataset scale. To avoid data leakage, we enforce synset-level separation across splits, so that no WordNet synset identifier is shared between the training, validation and test partitions. Each image pair is included in both presentation orders to reduce residual position bias. After preprocessing and augmentation the dataset contains 3536, 233 and 350 samples for training, validation and testing.

Architecture and optimisation. The model uses an OpenAI CLIP ViT-L/14 backbone with the last two vision and text transformer layers unfrozen, together with a projection head, optimised with a Bradley-Terry pairwise loss. Fine-tuning uses dropout 0.3, label smoothing 0.1, and a learning rate of 3×10^{-5}.



(a) The distribution of rewards across TTI models.



(b) Distribution of Tie and Both Bad rewards. Mann-Whitney U, $p < 0.001$

Figure 7: The distribution of rewards generated by fine-tuned CLIP model. Prompt does not include the definition.

Comparability. For the correlation analysis against human rankings, every generator is evaluated on the same set of test-set synsets, each producing one image per concept. Identical synset coverage across generators makes the reward and human preference statistics directly comparable.

Results without definitions Although a slight decrease in accuracy is observed, the magnitude of this reduction is relatively small, suggesting that the inclusion of definitions does not exert a substantial impact on overall performance. The correlation coefficients are marginally lower ($\rho = 0.92$ (95% CI: $[+0.875, +0.923]$) and accompanied by wider confidence intervals. As shown in Figure 7a FLUX and Playground again occupy the leading positions. Furthermore, the distinction between the Both Bad and Tie categories remains clearly preserved (Figure 7b).

D ELO Score Computation

This appendix gives the procedure behind the ELO scores reported in Section 5, so that it can be followed without consulting Chiang et al. (2024a).

Collecting comparisons. For every concept v in the evaluation set we draw two distinct systems A and B uniformly from the 12 under evaluation and show their images (x^A, x^B) side by side. Each system is equally likely to occupy either position, so no alternation of positions is needed. A judge, either one of our four assessors or GPT-4, returns one of four labels: *A wins*, *B wins*, *Tie*, or *Both Bad*. This gives 3370 judged pairs, roughly 600 involving each system.

Fitting the model. Only the decisive labels enter the fit; *Tie* and *Both Bad* are recorded but dropped. Writing $\pi_i > 0$ for the latent strength of system i, the Bradley-Terry model sets $P(i \succ j) = \pi_i/(\pi_i + \pi_j)$. With w_{ij} the number of times i was preferred over j, we maximise

$$\ell(\pi) = \sum_{i \neq j} w_{ij} \log \frac{\pi_i}{\pi_i + \pi_j} \tag{4}$$

with respect to π. The likelihood depends on π only through ratios, so the scale is fixed by one normalisation. We reuse the implementation of Chiang et al. (2024a) unchanged, which maps the fitted strengths onto the ELO scale as $1000 + 400 \log_{10} \pi_i$.

Confidence intervals. The 95% intervals in Figure 5 and Figure 19 are obtained by bootstrapping over the judged comparisons: we resample the set of pairs with replacement, refit the model on each resample, and report the empirical 2.5th and 97.5th percentiles of each system's score.

Judges. Human ELO uses four assessors with a background in computational linguistics and at least a bachelor's degree; their Spearman correlation is 0.8 ($p \leq 0.05$) on the images generated with definitions. GPT-4 ELO uses `gpt-4o-mini` with the prompt shown in Figure 8. Section 6 compares the two rankings.

E GPT-4 Prompts

We show the technical style prompt for GPT-4 in Figure 8 for more clarity on how images and user prompt were provided. We employed "gpt-4o-mini" version with API calls with images in high resolution. The prompt for "gpt-4o-mini" to generate definitions for TaxoLLaMA3.1 predictions is presented below.

```
Write a definition for the word/phrase in one sentence.

Example:
Word: caddle
Definition: act as a caddie and carry clubs for a player

Word: bichon
Definition:
```

```
Please act as an impartial judge and evaluate the quality of the images provided by two AI
image assistants to the user prompt displayed below.


You should choose the assistant that provides an image that follows the user's instructions
better and reflects the user's prompt main concept better.
Your evaluation should consider factors such as image-text alignment, relevance, accuracy,
depth, and fidelity (overall image quality).
Begin your evaluation by comparing the two images and provide a short explanation.

Avoid any position biases and ensure that the order in which the responses were presented do
not influence your decision.

Do not allow the size of the images to influence your evaluation.
Do not favor certain names of the assistants.
Be as objective as possible.

After providing your explanation, output your final verdict by strictly following this format:

"[[A]]" if assistant A is better, "[[B]]" if assistant B is better, "[[C]]" for a tie and "
[[D]]" if both images are bad.

[User Prompt]

inherited disorder (a genetic condition passed down from parents to their offspring, often
resulting in physical or mental health problems)

[Start of the first image]

<img_a>

[End of the first image]

[Start of the first image]

<img_b>

[End of the first image]
```

Figure 8: Full prompt example for evaluating text-to-image assistants.

F Analysis of Image Generation Errors

This analysis is made for generation without definitions.

All models struggle with depicting

a. abstract concepts;

b. nonfrequent and specific words ("orifice.n.01" with the lemma "rima");

c. notions of people with specific functional role ("holder.n.02" with the lemma "holder", for example).

Abstract concepts are handled with the following:

1. Text in images (although not all models succeed in writing);

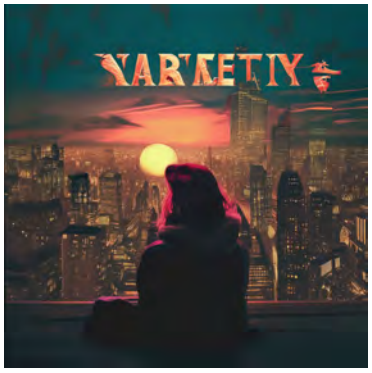
(a) Openjourney, prevention.n.01, prevention

(b) Openjourney, corporal_punishment.n.01, corporal punishment

(c) HDit, reform_movement.n.01, labor union

(d) SD3, murder.n.01, murder

(e) SD3, abstinence.n.02, abstinence



(f) PixArt, broadcast.n.01, headline

Figure 9: Examples of texts in images generated by Openjourney, HDiT, SD3, and PixArt.

2. Abstract images;

Other unwanted behaviors for the purposes of illustrating taxonomies include

1. Generating playing cards for the concepts (most seen in Openjourney, also present in SD1.5);

2. Abstract ornamental circles (also most found in Openjourne, and some in SD1.5).

3. Depicturing monsters when facing rare animal names (seen in Kandinsky3).

Most importantly, models struggle closer to the leaves of a taxonomy: they tend to create an image of a parent concept without necessary features of the child (see Figure 17).

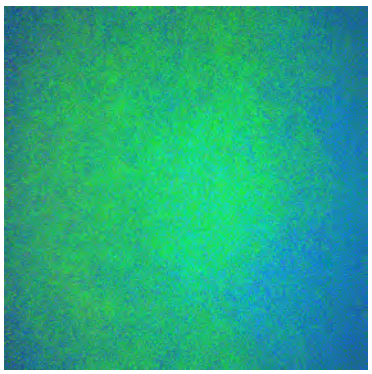
(a) PixArt, binomial.n.01, binomial

(b) PixArt, statement.n.01, statement

(c) HDiT, disrespect.n.01, obscenity

(d) SDXL, clearance.n.03, clearance

(e) HDiT, financial_gain.n.01, flow



(f) SD3, somesthesia.n.02, somesthesia

Figure 10: Examples of abstract images generated by PixArt, HDiT, SDXL, and SD3.

(a) Openjourney, sovereign.n.01, sovereign

(b) Openjourney, enthusiast.n.01, enthusiast

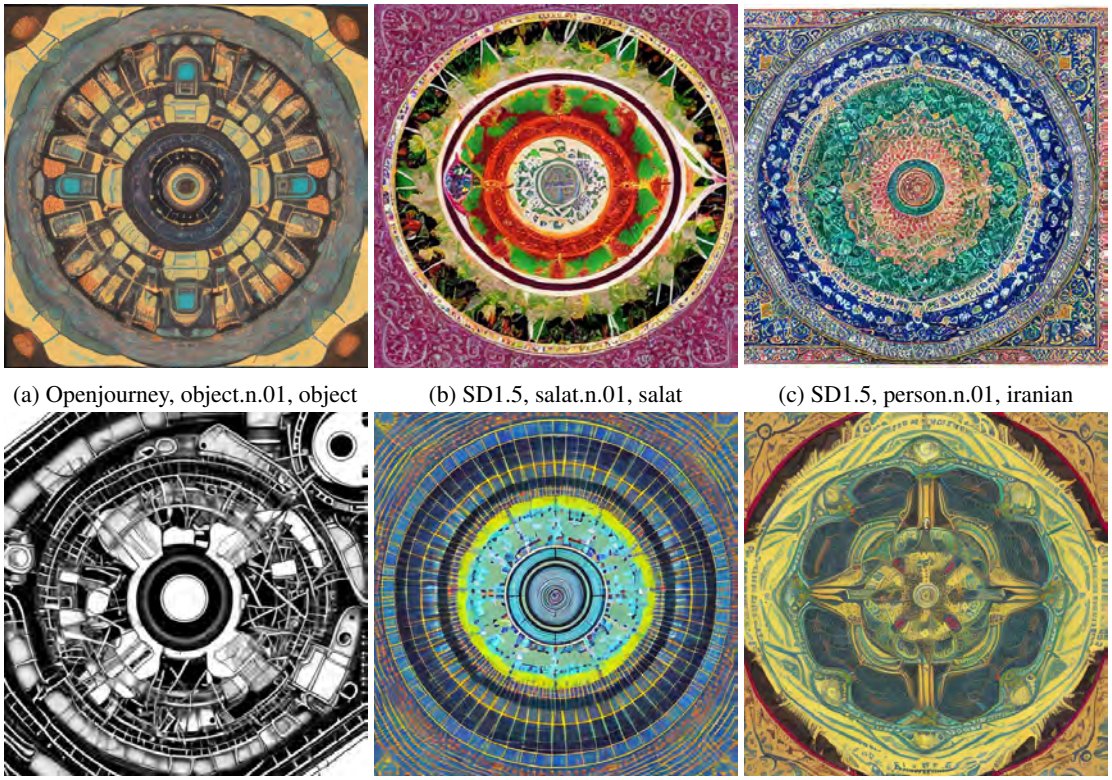
(c) Openjourney, policyholder.n.01, policyholder

(d) Openjourney, agreement.n.01, agreement

(e) Openjourney, ground-shaker.n.01, ground-shaker

(f) SD1.5, finisher.n.01, finisher

Figure 11: Examples of playing cards in Openjourney and SD1.5.



(a) Openjourney, object.n.01, object (b) SD1.5, salat.n.01, salat (c) SD1.5, person.n.01, iranian

(d) Openjourney, engineering.n.03, engineering (e) Openjourney, lingtpen.n.01, light pen (f) Openjourney, organization.n.01, baptist association

Figure 12: Examples of abstract ornamental circles generated by Openjourney and SD1.5.

(a) Kandinsky3, acanthocephalan.n.01, acanthocephalan (b) Kandinsky3, gelechiid.n.01, gelechiid

Figure 13: Examples of monsters generated for rare animal names by Kandinsky3.

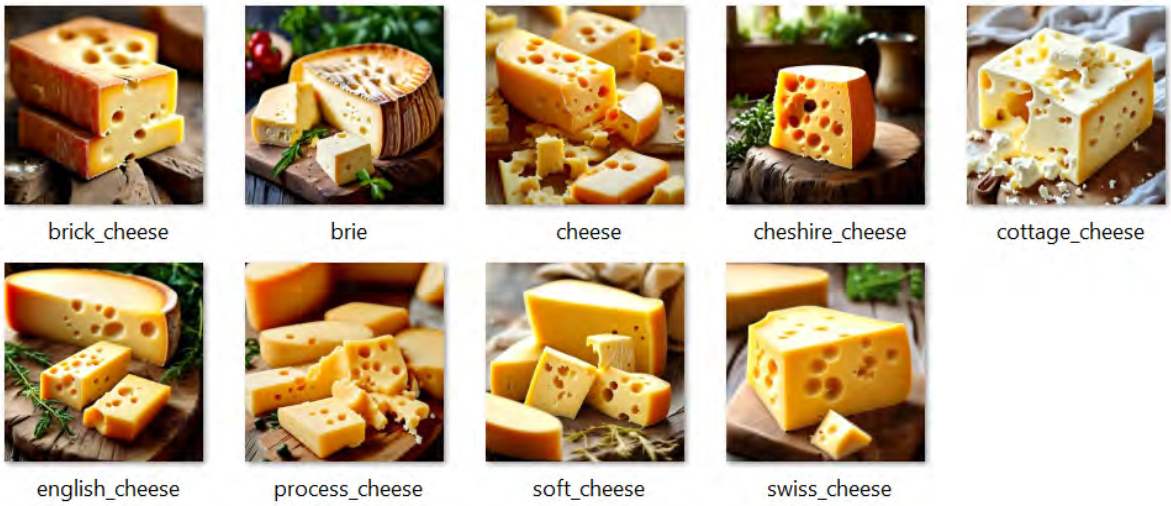


Figure 14: PixArt images for "cheese" and some of its hyponyms.

G Demonstration System Examples

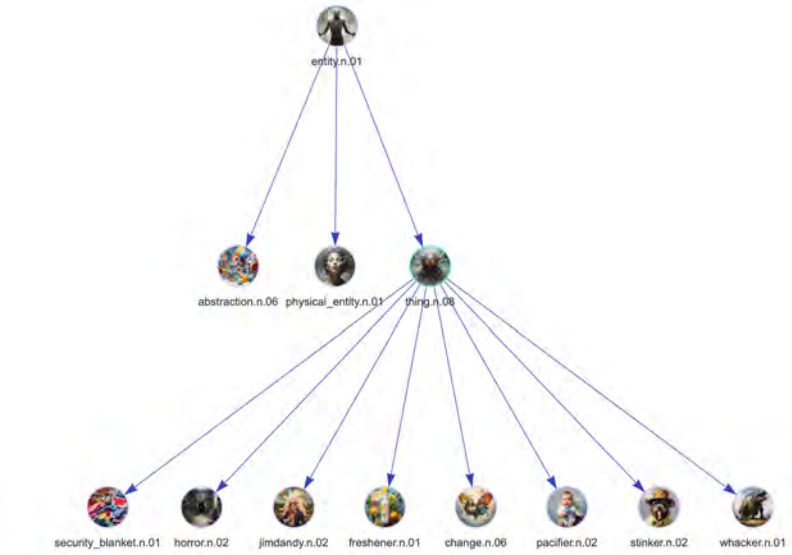


Figure 15: Subgraph starting from the root node “entity.n.01”. Images are generated with the best TTI model.

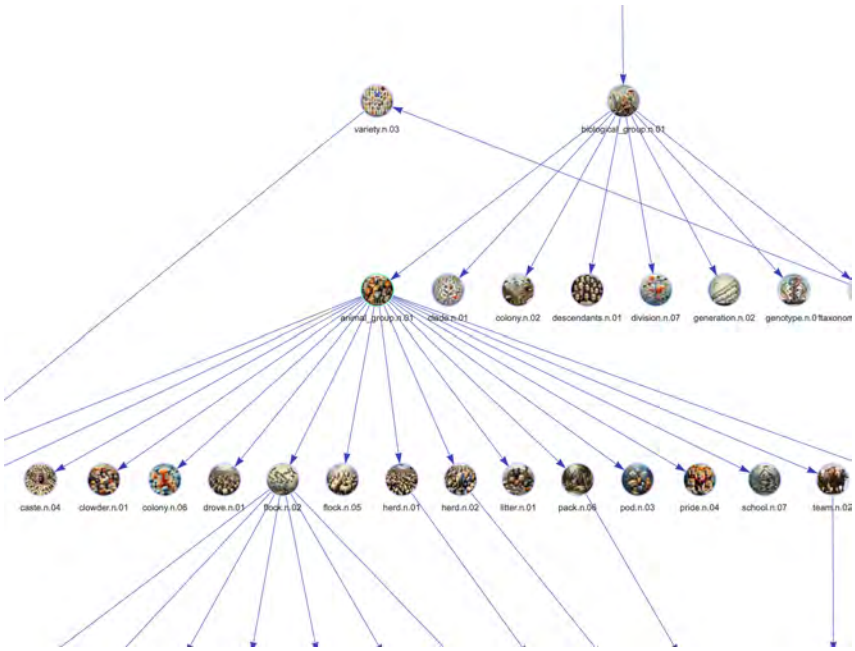


Figure 16: Subgraph starting from the node “biological_group.n.01”. Images are generated with the best TTI model.

biological_group.n.01

biological group

a group of plants or animals

pain.n.01

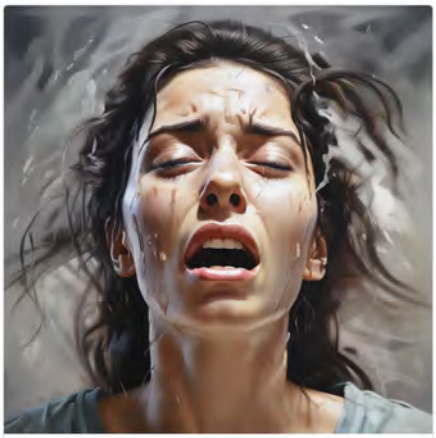
pain,hurting

a symptom of some physical hurt or disorder

feeling.n.01

feeling

the experiencing of affective and emotional states



emotion.n.01

emotion

any strong feeling

Figure 17: Node descriptions from the demonstration system with the generated image using the best-performing model.

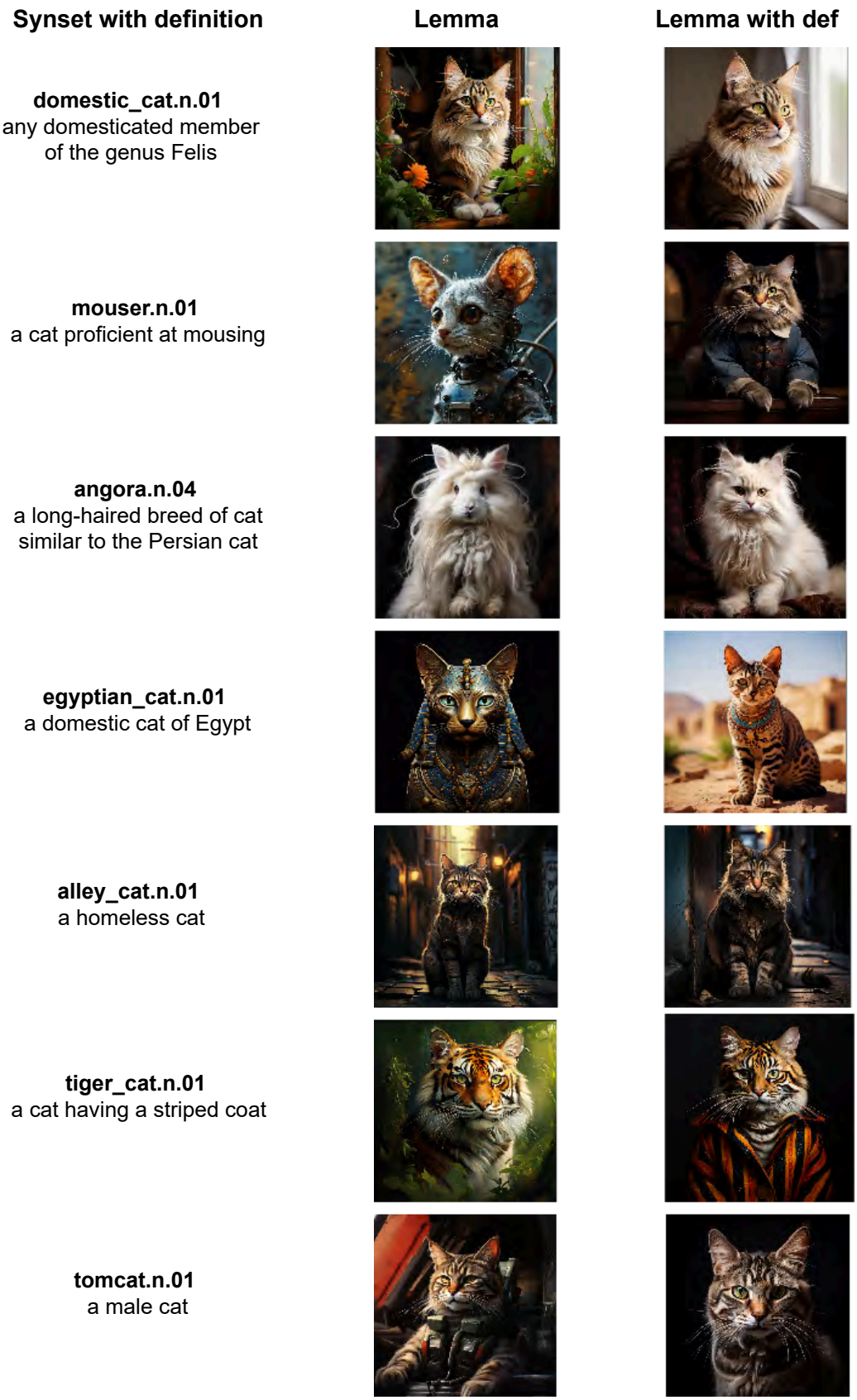


Figure 18: Examples of the generations for the 'domestic_cat.n.01' hyponyms using playground-v2.5-1024px-aesthetic model. The examples illustrate common confusion patterns without introducing an additional labeling step: mouser has features from "Mouser Electronics" and a mouse; Angora is mixed within a rabbit and has wool patterns; tiger cat either looks like a tiger or has a striped coat; tomcat is confused with aircraft and has made a cat a military pilot.

H Additional Figures

In this appendix, we include graphs for our evaluation, which results outlined in the main text.

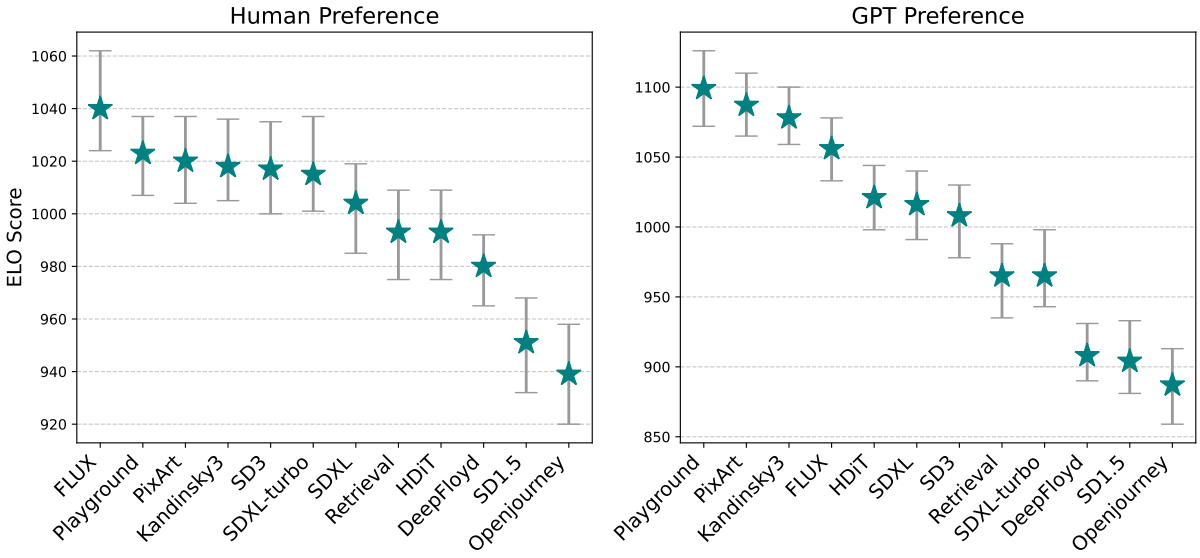


Figure 19: ELO scores for human and GPT-4 preferences. Prompt did not include the definition.

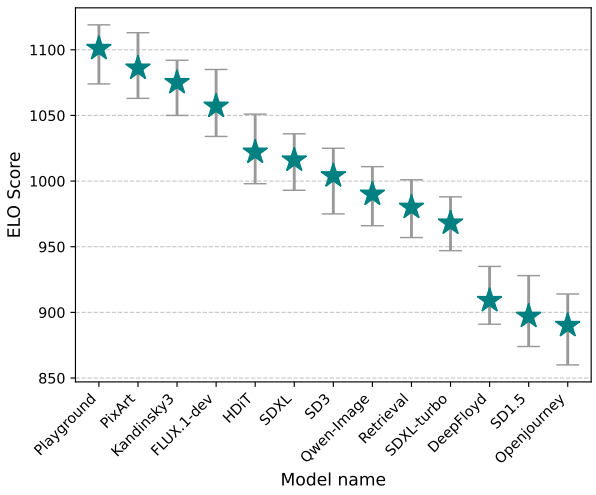


Figure 20: ELO scores for GPT-4 preferences with Qwen-Image model. Prompt did not include the definition.

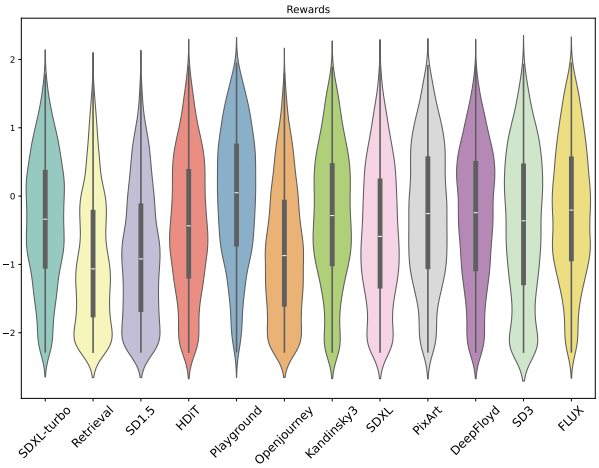


Figure 21: Distribution of rewards for each model, calculated with reward model described in Section 5

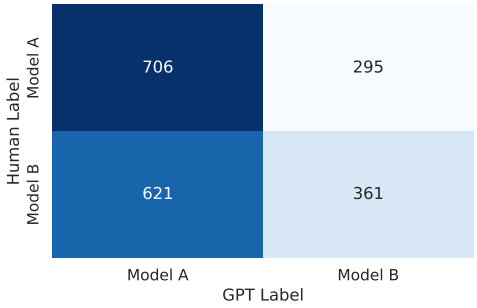


Figure 22: Confusion matrix for human and GPT preferences, excluding Tie labels to avoid distracting the analysis. GPT rarely assigns Ties, with fewer than 20 instances. The prompt included a definition.

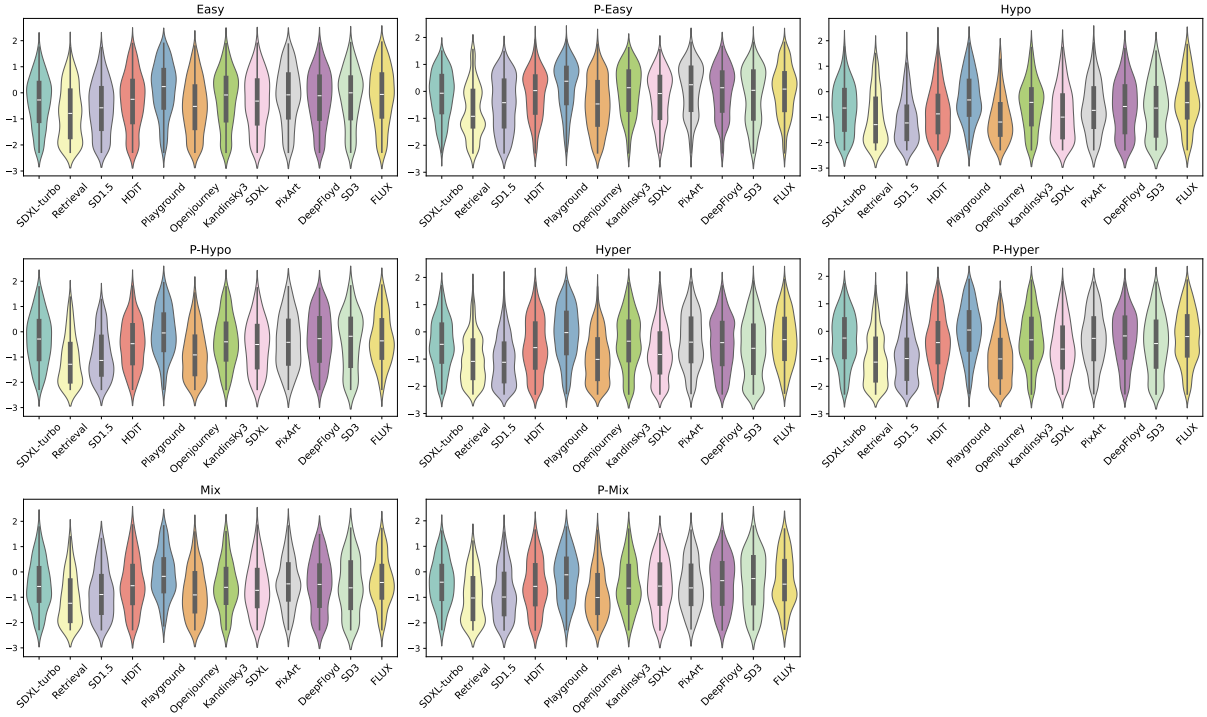


Figure 23: Distribution of rewards for each model across subsets, calculated with reward model described in Section 5.

I Additional Tables

Model	IS		FID	
	def	no_def	def	no_def
DeepFloyd	19.6	12.1	62	131
Kandinsky3	19.4	18.2	64	75
PixArt	19.8	18.8	73	73
Playground	20.9	20.0	71	74
Openjourney	15.4	13.9	68	73
SD1.5	18.0	17.5	59	60
SDXL-turbo	10.9	12.1	89	65
SD3	21.2	23.5	63	65
HDiT	18.2	20.2	67	85
SDXL	19.1	19.6	63	68
FLUX	20.9	22.0	68	72

Table 4: Comparing FID and IS for datasets with and without definition

Model	Ground Truth				Predicted			
	Hypo	Hyper	Mix	Easy	Hypo	Hyper	Mix	Easy
DeepFloyd	8.2	9.3	6.6	12.9	8.3	11.3	6.8	7.3
Kandinsky-3	7.4	10.3	6.6	12.8	7.8	11.7	7.0	8.4
PixArt	7.6	10.5	6.6	13.5	7.6	11.2	7.1	8.6
Playground	8.6	11.0	7.0	13.3	8.0	12.0	7.8	8.3
Openjourney	6.7	8.2	6.3	12.5	6.9	9.0	6.4	7.0
SD1.5	7.1	9.0	6.9	12.9	7.4	9.5	6.6	8.1
SDXL-turbo	5.2	6.7	4.7	9.9	5.3	6.8	5.1	6.1
SD3	8.4	9.5	7.2	13.3	8.1	12.0	7.3	8.8
HDiT	8.0	10.1	6.8	11.2	7.3	11.8	7.0	7.7
SDXL	7.6	10.4	7.1	12.8	7.5	11.4	7.3	8.1
FLUX	8.3	10.5	7.5	12.9	8.3	12.7	7.2	8.0
Retrieval	7.8	10.5	6.7	11.4	8.5	12.7	6.7	8.2

Table 5: Inception Score per subsets with definitions

Model	Ground Truth				Predicted			
	Hypo	Hyper	Mix	Easy	Hypo	Hyper	Mix	Easy
DeepFloyd	195	134	191	128	220	219	147	193
Kandinsky3	203	134	197	127	229	224	155	191
PixArt	203	143	191	134	229	230	163	203
Playground	198	143	188	130	234	225	163	196
Openjourney	200	145	197	127	225	228	159	198
SD1.5	192	135	192	125	229	221	154	186
SDXL-turbo	207	158	211	146	235	226	170	203
SD3	193	135	193	133	225	218	153	184
HDiT	201	141	187	126	226	225	160	198
SDXL	192	133	199	129	230	223	163	191
FLUX	163	132	186	139	165	115	182	171

Table 6: Fréchet Inception Distance Score per subsets with definitions

Model	CLIP-Score	Hypernym CLIP-Score	Cohyponym CLIP-Score	Specificity
HDiT	0.69	0.60	0.58	1.2
Retrieval	0.64	0.57	0.56	1.16
Kandinsky3	0.69	0.61	0.59	1.19
Openjourney	0.68	0.59	0.57	1.2
DeepFloyd	0.68	0.60	0.58	1.18
SDXL-turbo	**0.72**	**0.62**	**0.60**	1.23
Playground	0.70	0.60	0.58	1.22
SDXL	0.69	0.60	0.58	1.2
PixArt	0.68	0.60	0.58	1.19
SD3	0.70	0.60	0.58	1.21
SD1.5	0.69	0.59	0.57	**1.23**
FLUX	0.68	0.61	0.58	1.17

Table 7: Summary of CLIPscore Metrics Across Models

Model	Ground Truth				Predicted			
	Easy	Hypo	Hyper	Mix	Easy	Hypo	Hyper	Mix
HDiT	0.73	0.66	0.67	0.68	0.72	0.67	0.68	0.68
Retrieval	0.68	0.60	0.62	0.66	0.63	0.63	0.63	0.69
Kandinsky3	0.73	0.67	0.67	0.69	0.73	0.68	0.68	0.69
Openjourney	0.71	0.66	0.66	0.68	0.71	0.67	0.66	0.69
DeepFloyd	0.71	0.67	0.66	0.69	0.70	0.67	0.66	0.69
SDXL-turbo	0.76	0.71	0.71	0.73	0.75	0.72	0.71	0.74
Playground	0.74	0.67	0.68	0.70	0.72	0.69	0.68	0.70
SDXL	0.73	0.67	0.67	0.69	0.73	0.69	0.68	0.70
PixArt	0.72	0.66	0.67	0.69	0.72	0.67	0.67	0.68
SD3	0.73	0.68	0.67	0.70	0.72	0.69	0.68	0.70
SD1.5	0.73	0.68	0.67	0.70	0.72	0.69	0.68	0.70
FLUX	0.71	0.65	0.66	0.68	0.72	0.68	0.67	0.69

Table 8: Lemma CLIPScore Across Different Subsets

Model	Ground Truth				Predicted			
	Easy	Hypo	Hyper	Mix	P-Easy	P-Hypo	P-Hyper	P-Mix
HDiT	0.65	0.60	0.61	0.61	0.57	0.59	0.59	0.60
Retrieval	0.63	0.56	0.57	0.60	0.54	0.57	0.56	0.60
Kandinsky3	0.66	0.61	0.61	0.62	0.58	0.60	0.59	0.61
Openjourney	0.63	0.59	0.59	0.60	0.57	0.58	0.59	0.59
DeepFloyd	0.64	0.60	0.60	0.62	0.57	0.59	0.60	0.60
SDXL-turbo	0.67	0.62	0.62	0.63	0.59	0.61	0.60	0.61
Playground	0.66	0.60	0.61	0.61	0.57	0.59	0.59	0.60
SDXL	0.66	0.60	0.61	0.61	0.58	0.59	0.59	0.60
PixArt	0.65	0.60	0.60	0.62	0.57	0.59	0.59	0.61
SD3	0.66	0.61	0.61	0.62	0.57	0.60	0.60	0.60
SD1.5	0.64	0.60	0.60	0.61	0.57	0.58	0.60	0.59
FLUX	0.70	0.62	0.61	0.63	0.57	0.59	0.59	0.6

Table 9: Hypernym CLIPScore Across Different Subsets

Model	Ground Truth				Predicted			
	Easy	**Hypo**	**Hyper**	**Mix**	**P-Easy**	**P-Hypo**	**P-Hyper**	**P-Mix**
HDiT	0.61	0.59	0.58	0.60	0.54	0.57	0.58	0.58
Retrieval	0.59	0.55	0.55	0.59	0.51	0.56	0.55	0.58
Kandinsky3	0.61	0.59	0.59	0.61	0.55	0.58	0.59	0.59
Openjourney	0.59	0.58	0.57	0.59	0.54	0.57	0.57	0.57
DeepFloyd	0.61	0.58	0.58	0.61	0.55	0.57	0.58	0.59
SDXL-turbo	0.62	0.60	0.60	0.62	0.56	0.59	0.60	0.60
Playground	0.60	0.58	0.58	0.60	0.54	0.57	0.58	0.58
SDXL	0.61	0.58	0.58	0.60	0.55	0.58	0.58	0.60
PixArt	0.60	0.59	0.58	0.61	0.54	0.57	0.58	0.59
SD3	0.61	0.59	0.58	0.60	0.54	0.58	0.58	0.59
SD1.5	0.60	0.58	0.57	0.60	0.54	0.56	0.57	0.57
FLUX	0.63	0.59	0.59	0.61	0.54	0.57	0.57	0.58

Table 10: Cohyponym CLIPScore Across Different Subsets

Model	Ground Truth				Predicted			
	Easy	**Hypo**	**Hyper**	**Mix**	**P-Easy**	**P-Hypo**	**P-Hyper**	**P-Mix**
HDiT	1.21	1.14	1.15	1.16	1.34	1.18	1.18	1.18
Retrieval	1.17	1.11	1.13	1.14	1.24	1.15	1.15	1.18
Kandinsky3	1.21	1.14	1.15	1.16	1.32	1.18	1.19	1.18
Openjourney	1.22	1.16	1.15	1.17	1.32	1.18	1.18	1.21
DeepFloyd	1.18	1.16	1.14	1.15	1.29	1.16	1.16	1.18
SDXL-turbo	1.23	1.18	1.18	1.20	1.34	1.22	1.22	1.24
Playground	1.24	1.16	1.17	1.20	1.34	1.21	1.20	1.21
SDXL	1.22	1.15	1.15	1.18	1.33	1.20	1.19	1.20
PixArt	1.20	1.14	1.15	1.16	1.34	1.17	1.18	1.17
SD3	1.23	1.17	1.16	1.19	1.34	1.20	1.20	1.20
SD1.5	1.24	1.19	1.18	1.20	1.34	1.22	1.21	1.23
FLUX	1.14	1.10	1.12	1.13	1.32	1.18	1.18	1.20

Table 11: Specificity Scores Across Different Models and Subsets

J TTI Models Description

To generate the images, we employed eleven models and one retrieval approach, 12 systems in total.

J.1 U-Net-based models

Models based on the U-Net architecture:

- **SD-v1-5** (400M) (Rombach et al., 2022) is a SD-v1-2 fine-tuned on 595k steps at resolution 512x512 on "laion-aesthetics v2 5+" and 10% dropping of the text-conditioning to improve classifier-free guidance sampling.
- **SDXL** (6.6B) (Podell et al., 2024). Its U-Net is 3 times larger than in classical SD models, and an additional CLIP (Radford et al., 2021) text encoder further increases the parameter count.
- **SDXL Turbo** (3.5B) (Liu et al., 2024) is a distilled version of SDXL-1.0.
- **Kandinsky 3** (12B) (Arkhipkin et al., 2023). The sizes of U-Net and text encoders were significantly increased in comparison to the second generation.
- **Playground-v2-aesthetic** (2.6B) (Li et al., 2023) has the same architecture as SDXL, and is trained on a dataset from Midjourney[4].
- **Openjourney** (123M) (Prompthero, 2023) is also trained on Midjourney images.

J.2 Diffusion Transformers models

Diffusion Transformers (DiTs) models:

- **IF** (4.3B) (DeepFloyd.Lab, 2023). A modular system consisting of a frozen text encoder and three sequential pixel diffusion modules.
- **SD3** (2B) (Esser et al., 2024) is a Multimodal DiT (MMDiT). The authors used two CLIP encoders and T5 (Raffel et al., 2020) for combining visual and textual inputs.
- **PixArt-Sigma** (900M) (Chen et al., 2024b). The authors employed novel attention mechanism for the sake of efficiency and high-quality training data for 4K images.
- **Hunyuan-DiT** (1.5B) (Li et al., 2024) is a text-to-image diffusion transformer designed for fine-grained understanding of both English and Chinese, using a custom-built transformer structure and text encoder.
- **FLUX** (12B) (BlackForestLabs, 2024) is a rectified flow Transformer capable of generating images from text descriptions. It is based on a hybrid architecture of multimodal and parallel diffusion transformer blocks.

J.3 Retrieval

We retrieved images from Wikimedia Commons[5], following previous studies (Ferrada et al., 2018; Jones and Oyen, 2022). For 3370 total items, this process resulted in 1,790 unique images. For 20 concepts (32 dataset entities), no images were found. For 146 lemmas, the search returned images that had already been retrieved, likely due to the similarity of the concepts searched. We use the top-1 output from the main image search engine[6].

[4] https://www.midjourney.com

[5] https://commons.wikimedia.org/

[6] For the lemma "coin", the search URL is https://commons.wikimedia.org/w/index.php?search=coin&title=Special:MediaSearch&go=Go&type=image

K Metrics Definition

Let V be a finite set of concepts (lemmas), where each $v \in V$ represents a semantic category. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measurable space representing the image domain, where $\mathcal{X}$ is the set of all possible images, $\mathcal{A}$ is a σ-algebra of measurable subsets of $\mathcal{X}$, and μ is a base measure.

For each concept $v \in V$, we assume a probability measure $P(\cdot \mid v)$ on $(\mathcal{X}, \mathcal{A})$ such that $P(X \in A \mid v)$ represents the probability that an image X generated under the concept v lies in the measurable set A. In other words, each concept v defines a probability distribution over the image space $\mathcal{X}$.

K.1 Lemma Similarity

Definition (Lemma Similarity)**.** *Given a concept $v \in V$ and an image $x \in \mathcal{X}$, the* Lemma Similarity *is defined as:*

$$S_{\textit{lemma}}(v, x) := P(X = x \mid v).$$

This measures the likelihood of observing the image x under the assumption that the concept v is the generating source. According to Theorem 1, it also reflects the likelihood of the concept v given the image x, offering a principled way to assess how well an image aligns with the semantic properties of a concept. Maximizing this metric implies a stronger alignment between the concept and the generated image.

Theorem 1. *Let V be a finite set of concepts, and suppose the prior distribution is uniform $P(V = v) = \frac{1}{|V|}\, \forall v \in V$. Then, $\forall x \in X\, \forall v \in V$ $\arg\max_{i \in V} S_{\textit{lemma}}(v, x) \propto \arg\max_{i \in V} P(V = v \mid X = x)$.*

Proof. By Bayes' rule, the posterior probability of concept i given an image x is:

$$P(V = i \mid X = x) = \frac{P(X = x \mid i)P(V = i)}{\sum_{v \in V} P(X = x \mid v)P(V = v)} = \frac{P(X = x \mid i) \cdot \frac{1}{|V|}}{\sum_{v \in V} P(X = x \mid v) \cdot \frac{1}{|V|}} = \frac{P(X = x \mid i)}{\sum_{v \in V} P(X = x \mid v)} \propto P(X = x \mid i) \tag{5}$$

To find the concept i that maximizes the posterior $P(V = i \mid X = x)$, we write:

$$\hat{i} = \arg\max_{i \in V} P(V = i \mid X = x) = \arg\max_{i \in V} P(X = x \mid i) = \arg\max_{i \in V} S_{\text{lemma}}(i, x).$$

Thus, under a uniform prior, maximizing the posterior probability is equivalent to maximizing the Lemma Similarity □

K.2 Hypernym Similarity & Cohyponym Similarity

Definition (Hypernym Similarity)**.** *Let $A(i) \subseteq V$ be the set of hypernyms of a concept $i \in V$. For a given image $x \in \mathcal{X}$, we define the* Hypernym Similarity *as:*

$$S_{\textit{hyper}}(i, x) := P(X = x \mid A(i)) = \frac{1}{|A(i)|} \sum_{h \in A(i)} P(X = x \mid h).$$

Definition (Cohyponym Similarity)**.** *Let $C(i) \subseteq V$ be the set of cohyponyms of a concept $i \in V$. For a given image $x \in \mathcal{X}$, we define the* Cohyponym Similarity *as:*

$$S_{\textit{cohyponym}}(i, x) := P(X = x \mid C(i)) = \frac{1}{|C(i)|} \sum_{\textit{ch} \in C(i)} P(X = x \mid \textit{ch}).$$

Hypernym Similarity and *Cohyponym Similarity* represent the likelihood of observing x under the average distribution of its ancestor and cohyponyms concepts respectively. Intuitively, they measure how well the image x fits into the neighboring concepts either broader, more general semantic category represented by the ancestors of i or similar, slightly different concepts from the same ancestor of i. According further to Theorem 2, maximizing those similarities is proportional to minimizing distance between image space conditioned on a concept and image space conditioned on specific neighbor space, therefore better reflecting neighbors semantic properties and covering tree structure.

Theorem 2. *With large enough $S_{lemma}(i,x)$ to properly represent our concept,* $\max S_{hyper}(i,x)$ *and* $\max S_{cohyponym}(i,x) \propto \min_{P(X|A(i))} D_{KL}\big(P(X \mid i) \,\|\, P(X \mid A(i))\big)$.

Proof. The proof will be based for the ancestor case, however proving for cohyponyms is similar with only change of ancestor set to cohyponyms set.

Consider the KL divergence:

$$D_{\mathrm{KL}}\big(P(X \mid i) \,\|\, P(X \mid A(i))\big) = \sum_{x \in \mathcal{X}} P(X = x \mid i) \log \frac{P(X = x \mid i)}{P(X = x \mid A(i))}.$$

$$\arg \min_{P(X|A(i))} D_{\mathrm{KL}}\big(P(X \mid i) \,\|\, P(X \mid A(i))\big) \implies \frac{P(X = x \mid i)}{P(X = x \mid A(i))} \approx 1 \quad \forall x.$$

As $P(X = x \mid i)$ is fixed in precise setting, achieving $\frac{P(X=x|i)}{P(X=x|A(i))} \approx 1$ requires increasing $P(X = x \mid A(i))$ subject to large enough $P(X = x \mid i)$. Since $S_{\text{hyper}}(i,x) = P(X = x \mid A(i))$, increasing $S_{\text{hyper}}(i,x)$ for the most probable x under i reduces the KL divergence. □

K.3 Specificity

Definition (Specificity)**.** *Let $C(i)$ be the set of cohyponyms of a concept $i \in V$. Define:*

$$P(X = x \mid C(i)) := \frac{1}{|C(i)|} \sum_{c \in C(i)} P(X = x \mid c).$$

The Specificity *of an image $x \in \mathcal{X}$ with respect to a concept $i \in V$ is:*

$$\textit{Spec}(i,x) := \frac{P(X = x \mid i)}{P(X = x \mid C(i))}.$$

Specificity measures how much more likely it is that x was generated by a concept i compared to one of its cohyponyms. A high Specificity value indicates that the probability of x under i is significantly larger than under $C(i)$, the distribution over its cohyponyms. According to Theorems 3 and 4, Specificity highlights the uniqueness of the node representation by maximizing the distance to the cohyponyms nodes distribution and increasing the mutual information between the concept and image spaces. However, this metric relies on a high *Lemma Similarity* value; otherwise, the reflected uniqueness may be misleading due to poor alignment with the target node's distribution.

Theorem 3.

$$\max \textit{Spec}(i,x) \propto \max D_{KL}(P(X \mid i) \| P(X \mid C(i)))$$

Proof.

$$\mathrm{Spec}(i,x) = \frac{P(X = x \mid i)}{P(X = x \mid C(i))}.$$

$$\log(\mathrm{Spec}(i,x)) = \log \frac{P(X = x \mid i)}{P(X = x \mid C(i))}.$$

$$D_{\mathrm{KL}}(P(X \mid i) \| P(X \mid C(i))) = \sum_{x} P(X = x \mid i) \log \frac{P(X = x \mid i)}{P(X = x \mid C(i))}.$$

For fixed $P(X = x \mid i)$, increasing $\frac{P(X=x|i)}{P(X=x|C(i))}$ for all x (i.e., maximizing $\mathrm{Spec}(i,x)$) increases each $\log \frac{P(X=x|i)}{P(X=x|C(i))}$ term. Thus, the sum $D_{\mathrm{KL}}(P(X \mid i) \| P(X \mid C(i)))$ is maximized. □

Theorem 4. *Let V and X be random variables over concepts and images.* $\max \textit{Spec}(i,x) \forall v \in V, x \in X \propto \max I(V;X)$.

Proof.

$$I(V;X) = \sum_{v,x} P(V=v, X=x) \log \frac{P(X=x \mid V=v)}{P(X=x)}.$$

For $v = i$:

$$\frac{P(X=x \mid i)}{P(X=x)} = \frac{P(X=x \mid i)}{\sum_{v'} P(V=v')P(X=x \mid v')}.$$

Consider the uniform prior ($P(V=v') = \frac{1}{|V|}$):

$$P(X=x) = \frac{1}{|V|} \sum_{v'} P(X=x \mid v'), \quad P(X=x \mid C(i)) = \frac{1}{|C(i)|} \sum_{c \in C(i)} P(X=x \mid c).$$

$$\text{Spec}(i,x) = \frac{P(X=x \mid i)}{P(X=x \mid C(i))}.$$

Increasing $\text{Spec}(i,x) \propto \frac{P(X=x|i)}{P(X=x)}$, raising each term $P(V=i, X=x) \log \frac{P(X=x|i)}{P(X=x)}$, thus increasing $I(V;X)$. □