Multilingual Conceptual Coverage in Text-to-Image Models

Michael Saxon
University of California, Santa Barbara
saxon@ucsb.edu

William Yang Wang
University of California, Santa Barbara
william@cs.ucsb.edu

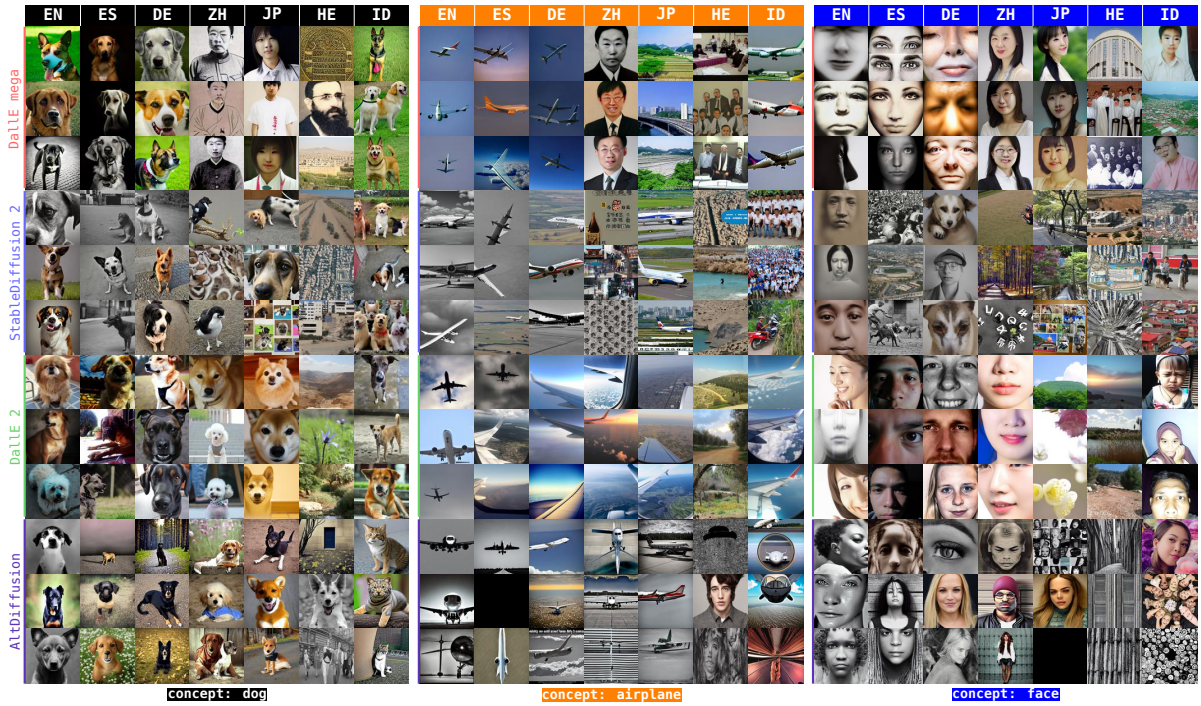


Figure 1: A selection of images generated by DALLE-mega, Stable Diffusion 2, DALLE-2, and AltDiffusion, illustrating their *conceptual coverage* of "**dog**," "**airplane**," and "**face**" across English, Spanish, German, Chinese (simplified), Japanese, Hebrew, and Indonesian. Coverage of the concepts varies considerably across model and language, and can be observed in the consistency and correctness of images generated under simple prompts.

Abstract

We propose "**Co**nceptual **Co**verage **Acro**ss **La**nguages" (***CoCo-CroLa***), a technique for benchmarking the degree to which any generative text-to-image system provides multilingual parity to its training language in terms of tangible nouns. For each model we can assess "conceptual coverage" of a given target language relative to a source language by comparing the population of images generated for a series of tangible nouns in the source language to the population of images generated for each noun under translation in the target language. This technique allows us to estimate how well-suited a model is to a target language as well as identify model-specific weaknesses, spurious correlations, and biases without a-priori assumptions. We demonstrate how it can be used to benchmark T2I models in terms of multilinguality, and how despite its simplicity it is a good proxy for impressive generalization.

1 Introduction

Neural text-to-image models convert plain text prompts into images (Mansimov et al., 2015; Reed et al., 2016) using internal representations reflective of the training data population. Advancements in conditional language modeling (Lewis et al., 2019), variational autoencoders (Kingma and Welling, 2013), GANs (Goodfellow et al., 2020), multimodal representations (Radford et al., 2021), and latent diffusion models (Rombach et al., 2022) have led to sophisticated text-to-image systems.

These models exhibit impressive *semantic generalization capabilities*, enabling them to generate coherent, visually-appealing images containing novel combinations of objects, scenarios, and styles (Ramesh et al., 2021; Saharia et al., 2022). They have semantic latent spaces (Kwon et al., 2022) by grounding words to their associated visuals (Hutchinson et al., 2022). However, **character-**

Figure 2: We hypothesize that a model's ability to generate creative, compositional images depicting tangible concepts (e.g., astronaut, horse, soup, bear) is predicated on its ability to generate simple images of the concepts alone. Samples from Ramesh et al. (2022).

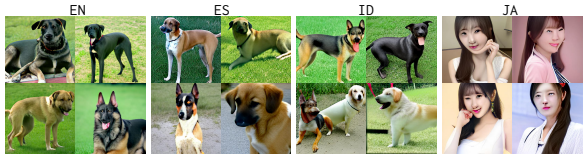


Figure 3: Although DALL-E mini (Dayma et al., 2021) is ostensibly trained only on English data, when elicited with "big dog" in Spanish, Indonesian, and Japanese it generalizes the "dog" concept to ES and ID, while exhibiting an offensive concept-level collision in JA.

izing the limits of these systems' capabilities is a challenge. They are composed of elements trained on incomprehensibly large (Prabhu and Birhane, 2020; Jia et al., 2021), web-scale data (Gao et al., 2020; Schuhmann et al., 2021), hindering training-data-centric model analysis (Mitchell et al., 2019; Gebru et al., 2021) to address this problem.

Demonstrations of novel T2I model capabilities tend to rely on the subjective impressiveness of their ability to generalize to complex, novel prompts (Figure 2). Unfortunately, the space of creative prompts is in principle infinite. However, we observe that **impressive creative prompts are composed of known, tangible concepts**.

Can we directly evaluate a model's knowledge of these tangible concepts as a partial proxy for its capability to generalize to creative novel prompts? Perhaps. But finding a diverse set of significant failure cases of basic concept knowledge for theses models is challenging—in their training language.

We observe that when prompted with simple requests for specific tangible concepts in a constrained style, T2I models can sometimes generate consistent and semantically-correct images in languages for which they received limited training (Figure 1, Figure 3). We refer to this capacity as *language-concept possession* by said model. At scale, we can assess the language-concept possession for a diverse array concepts and languages in a model to attempt to describe its overall multilingual generalization capability. We refer to the degree of this capability as the model's multilingual *conceptual coverage*. In this work we:

1. Introduce objective measures of *multilingual conceptual coverage* in T2I models that compare images generated from equivalent prompts under translation (Figure 4).

2. Release ***CoCo-CroLa***, a benchmark set for conceptual coverage testing for 193 tangible concepts across English, Spanish, German, Chinese, Japanese, Hebrew, and Indonesian.

3. Validate the utility of *conceptual coverage analysis* of T2I models with a pilot study providing evidence that generalization to complex, creative prompts is predicated on concept possession.

Our benchmark enables fine-grained concept-level model analysis, identification of novel failure modes, and will guide future work in increasing the performance, explainability, and linguistic parity of text-to-image models.

2 Motivation & Related Work

This work is an attempt to produce a scalable technique for characterizing models in terms of conceptual coverage across multiple languages, with minimal assumptions about the concepts or models themselves. In this section we lay out our motivations alongside relevant related work.

Benchmarks enabling model comparability have been a driving force in the development of pretrained language models (LM) (Devlin et al., 2018). For classification and regression tasks, evaluation under fine-tuning (Howard and Ruder, 2018; Karpukhin et al., 2020) is a straightforward and practical proxy for pretrained LM quality (e.g., for encoder-only transformer networks (Liu et al., 2019)) (Dodge et al., 2020). For these classification models, higher performance on benchmark datasets (Lai et al., 2017; Rajpurkar et al., 2018; Wang et al., 2019) became the primary target of LM advancement. However, other important qualities in models including degree of social biases (Sheng et al., 2019) and robustness (Clark et al., 2019) arising from biases in training data (Saxon et al., 2022) can only be captured by more sophisticated benchmarks that go beyond simple accuracy (Cho et al., 2021). CheckList represented an influential move in this direction by benchmarking

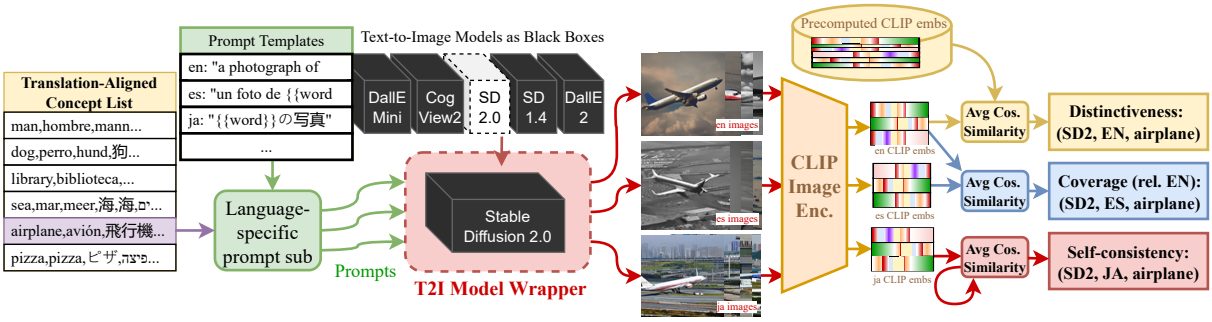


Figure 4: ***CoCo-CroLa*** assesses the cross-lingual coverage of a concept in a model by plugging all the term translations into prompt templates, generating a set of images from a model under test, extracting their corresponding CLIP embeddings, and computing concept-level **distinctiveness**, **coverage**, and **self-consistency** for the concept with respect to each language. (Anonymized demo available at `conceptualcoverage.github.io`)

model performance through *behavioral analysis under perturbed elicitation* (Ribeiro et al., 2020).

In contrast, generative large language models (LLMs) such as GPT-3 (Brown et al., 2020) have a broader range of outputs, use-cases, and capabilities, making evaluation more difficult. For many text-generative tasks such as summarization and creative text generation, the crucial desired quality is subjective, and challenging to evaluate (Xu et al., 2022). However, as these LLMs operate in a text-only domain, existing supervised tasks could be ported to few-shot or zero-shot evaluations of LLM capabilities (Srivastava et al., 2022). While performance on these benchmarks isn't directly indicative of the impressive generative performance and generalization capabilities, they are a means to measure improvement (Suzgun et al., 2022).

Text-to-image models are even more difficult to evaluate than LLMs. Unlike in LLMs, there aren't ready-made evaluation tasks that can be ported over. For example, while GPT-3 was introduced with impressive and sometimes SOTA performance on zero-shot generalization to a suite of classification tasks, the T2I model DALL-E 2 was primarily introduced with human opinion scores and cool demo images (Ramesh et al., 2022).

Multilingual conceptual coverage is a high-variation T2I model performance setting. (Figure 1) Perhaps more importantly, it has immediate value, as work on improving T2I model multilinguality has has been proposed, but hampered by a lack of evaluation metrics.

Chen et al. (2022) introduce AltCLIP and AltDiffusion, models produced by performing multilingual contrastive learning on a CLIP checkpoint for an array of non-English languages including Japanese, Chinese, and Korean. Without an objective evaluation benchmark, they can only demonstrate their improvement through human evaluation of impressive but arbitrary examples. ***CoCo-CroLa*** improves this state of affairs by enabling CheckList-like direct comparison of techniques for reducing *multilingual conceptual coverage* disparities as an objective, capabilities-based benchmark.

Excitingly, we find that **conceptual coverage is upstream of the impressive T2I model creativity** that model developers and end-users are fundamentally interested in. This means that not only is ***CoCo-CroLa*** an objective evaluation of T2I system capabilities, it is also a **proxy measure for the deeper semantic generalization capabilities we are interested in enhancing** in second languages, as we demonstrate in subsection 5.6.

3 Definitions & Formulations

We define a multilingual *concept* over languages L as a set of words in each language carrying the same meaning and analoguous colloquial use. We refer to the equivalent translation of concept c_k in language ℓ as $c_{k,\ell}$.

Given a set of concepts C, test language ℓ, a *minimal eliciting prompt*[1] MP_ℓ, text-to-image model f, and a desired number of images-per-concept n, we sample $n|C||L|$ images $I_{c_k,\ell,i}$, where

$$I_{c_k,\ell,i} \sim f(\text{MP}_\ell(c_{k,\ell})) \quad (1)$$

For every concept word in the language ℓ $c_k \in C$.

Given an image feature extractor F, some similarity function $\mathcal{L}(\cdot,\cdot)$, we assess whether f *pos-*

[1]We define a minimal eliciting prompt as a short sentence with a slot for concept work insertion, intended to enforce style consistency without interfering with the concept.

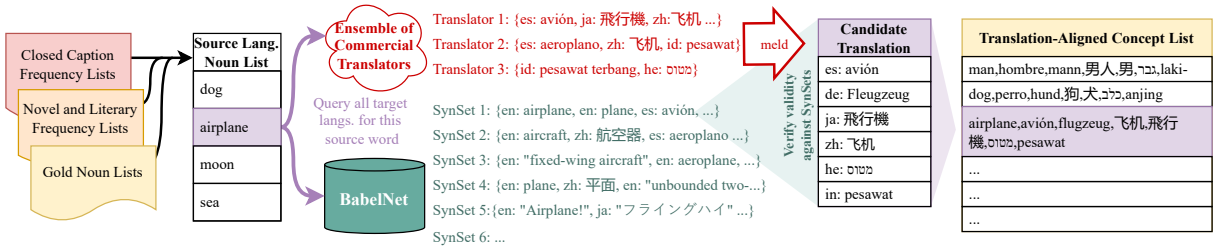


Figure 5: A diagram of our approach for producing the aligned noun concept list across the target language set using an ensemble of cloud translation services and BabelNet. Full description of this method in Appendix A.

sesses concept $c_{k,\ell}$ in the test language if using the following metrics from the concept-image set $\{I_{c_{k,\ell},i}\}_{i=0}^{n}$ (***CoCo-CroLa*** scores in Figure 4):

Distinctiveness. The images are *distinct* if they tend to not resemble the population of images generated for other concepts in the target language.

Formally, we compute the distinctiveness score $\mathrm{Dt}(f, \ell, c_k)$ relative to m images sampled from other concepts in C:

$$\mathrm{Dt} = \frac{1}{mn} \sum_{j=0}^{m} \sum_{i=0}^{n} \mathcal{L}(F(I_{c_k,\ell,i}), F(I_{c_r,\ell,s})), \quad (2)$$

$$c_{r,\ell} \sim C \setminus c_{k,\ell}, \quad s \sim U\{0, n\} \quad (3)$$

Self-consistency. The images are *self-consistent* if they tend to resemble each other as a set.

Formally, we compute the self-consistency score $\mathrm{Sc}(f, \ell, c_k)$ as:

$$\mathrm{Sc} = \frac{1}{n^2 - n} \sum_{j=0}^{n} \left(\sum_{i=0}^{n} \mathcal{L}(F(I_{c_k,\ell,i}), F(I_{c_k,\ell,j})) - 1 \right) \quad (4)$$

We subtract 1 from each step in the numerator and n from the denominator so that identical matches generated image to itself.

Correctness. The images are *correct* if they faithfully depict the object being described.

Rather than assess this using a classification model (hindering generality depending on the pretrained classifier), we use faithfulness relative to a source language ℓs, cross consistency $\mathrm{Xc}(f, \ell, c_k, \ell s)$ as a proxy:

$$\mathrm{Xc} = \frac{1}{n^2} \sum_{j=0}^{n} \sum_{i=0}^{n} \mathcal{L}(F(I_{c_k,\ell,i}), F(I_{c_k,\ell s,j})) \quad (5)$$

Additionally, we use the average text-image similarity score of the English concept text against the set of generated images, for a CLIP image encoder F and text encoder F_t, Wc:

$$\mathrm{Wc} = \frac{1}{n} \sum_{i=0}^{n} F_t(c_{k,\ell_s}) \cdot F(I_{c_k,\ell,i}) \quad (6)$$

4 Approach

We compute distinctiveness, self-consistency, and correctness scores across English, Spanish, German, Chinese (Simplified), Japanese, Hebrew, and Indonesian on the models listed in Table 1.

We use a CLIP (Radford et al., 2021) checkpoint from HuggingFace[2] as our semantic visual feature extractor F, and cosine similarity as our similarity function $\mathcal{L}(\cdot, \cdot)$. We collect a translation-aligned concept list C using techniques described in subsection 4.1 and depicted in Figure 5. We release our list generation code, testing code, feature extraction code, and final concept list as ***CoCo-CroLa*** v0.1[3].

4.1 Translation-aligned concept set collection

We implement an approach to automatically produce an aligned multilingual concept list, where meaning, colloquial usage, and connotations are preserved as well as possible. We identify *tangible nouns* describing physical objects, animals, people, and natural phenomena as a class of concepts that are both straightforward to evaluate and tend toward relative ubiquity in presence as words across languages and cultures.

Automated production is desirable for this task, as it enables the straightforward addition of new languages to the benchmark. To minimize translation errors utilize both a large knowledge graph of terminology and an ensemble of commercial machine translation systems to produce an aligned concept list (Figure 5). We accept a modest rate of mistranslations as the price of convenience and scale,

[2] HF:openai/clip-vit-base-patch32.

[3] Anon. demo @ conceptualcoverage.github.io

Model	Authors (Year)	Repository	Training Language
DALL-E Mini	Dayma et al. (2021)	`github:borisdayma/dalle-mini`	EN
DALL-E Mega	⎪	⎪	⎪
CogView 2	Ding et al. (2021)	`github:THUDM/CogView`	ZH
StableDiffusion 1.1	Rombach et al. (2022)	`HF:CompVis/stable-diffusion-v1-1`	EN
StableDiffusion 1.2	⎪	`HF:CompVis/stable-diffusion-v1-2`	⎪
StableDiffusion 1.4	⎪	`HF:CompVis/stable-diffusion-v1-4`	No language filter
StableDiffusion 2	⎪	`HF:stabilityai/stable-diffusion-2`	⎪
DALL-E 2	Ramesh et al. (2022)	`openai.com/dall-e-2/` (no checkpoints)	No language filter
AltDiffusion m9	Chen et al. (2022)	`HF:BAAI/AltDiffusion-m9`	EN, ES, FR, IT, RU, ZH, JA, KO

Table 1: The set of text-to-image models we evaluated with ***CoCo-CroLa*** v0.1. Some monolingual models may integrate pretrained elements such as CLIP checkpoints that have been trained on multilingual data.

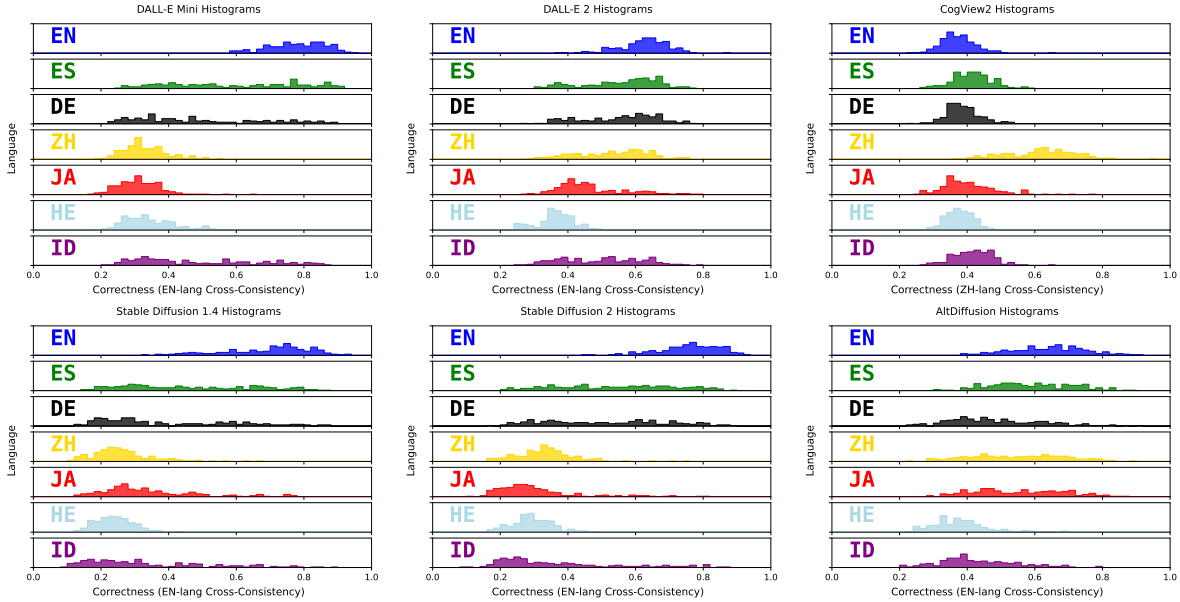


Figure 6: Histograms of the distribution of **correctness** cross-consistency (Xc) for each test language for six assessed models. Rightward probability mass reflects better conceptual coverage.

and analyze some examples of mistranslations, translation collisions, and their effects. Full details for our translation pipeline are in Appendix A.

4.2 Making minimal eliciting prompts

As discussed in section 3, an ideal prompt template would enforce stylistic consistency in the generated outputs without introducing biases that interfere with the demonstration of concept possession. Following Bianchi et al. (2022) we build simple prompts of the form, "a photograph of _____", which we manually translate into target languages. This simple template-filling approach will introduce grammatical errors for some languages. We briefly investigate if this matters in Appendix B.

4.3 Applying the metrics for analysis

We assess Dt, Sc, Xc, and Wc for each (concept, language) pair for each model. Using these we compare models and assess the validity of conceptual coverage as a proxy for generalization.

5 Findings

Figure 6 shows histograms for the distributions of the cross-consistency correctness proxy score Xc for each concept, relative to the training language (either English or Chinese) for DALL-E Mini, DALL-E 2, CogView 2, Stable Diffusion 1.4, Stable Diffusion 2, and AltDiffusion across the seven test languages. This plot clearly depicts that for the primarily English-trained models (DALL-E Mini, Stable Diffusion 1.4, Stable Diffusion 2), English-language performance (a high-mean distribution of high-EN-EN consistency concepts) is considerably better than the other languages. Similarly, for CogView2, trained on Chinese, the Chinese distribution of ZH-ZH scores is considerably better than the others, which do equally bad.

DALL-E 2 recieved open-ended multilingual training, and exhibits more consistent acceptable performance across the European and East Asian languages being tested. AltDiffusion, which has

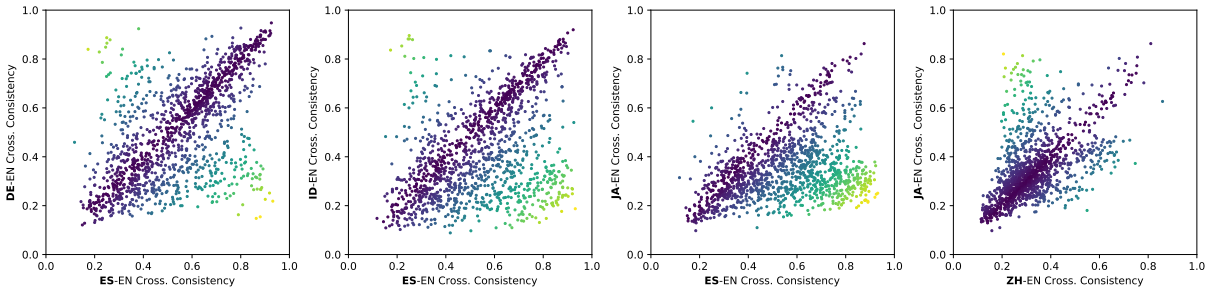
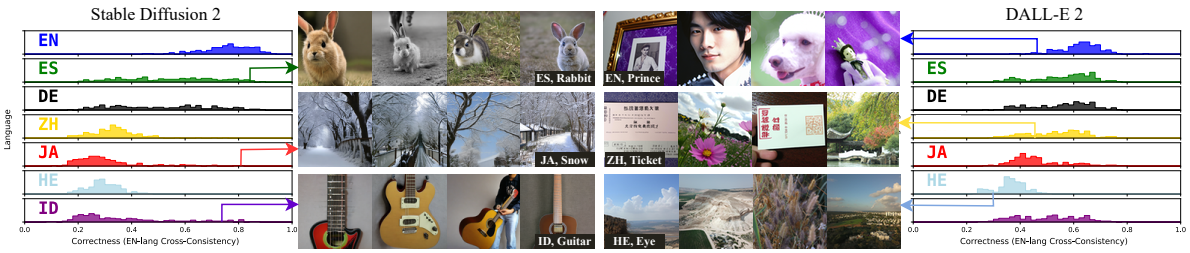


Figure 7: The **correctness** score for every (concept, model) pair for (right to left) ES vs DE, ES vs ID, ES vs JA, and ES vs JA. Languages sharing scripts (ES/DE/ID and JA/ZH) are more correlated than those that don't (ES/JA).



(a) High-coverage concepts (b) Low-coverage concepts

Figure 8: We automatically identify (a) high-coverage concepts in Stable Diffusion 2 (ES, rabbit), (JA, snow), (ID, guitar) and (b) low-coverage concepts in DALL-E 2 (EN, prince), (ZH, ticket), (HE, eye) using **correctness** Xc.

had its CLIP text encoder contrastively trained against multilingual representations on 9 languages (including ES, DE, ZH, and JA) exhibits higher performance on its training languages than its non-training languages (HE and ID).

Correctness distributions for Spanish, German, and Indonesian look roughly similar (in terms of mean and variance) for all models but AltDiffusion. This is particularly interesting because they are the three non-English languages that also use the Latin alphabet. Figure 7 compares the **correctness** Xc score for every concept, in every model, across pairs of languages that fully or partially share scripts (ES, DE, ID), (ZH, JA) and two languages that don't (JA, ES). Across pairs of languages that share scripts, there is a high correlation between possession of a given concept in one language and the other. A consistent trend across all models was poor performance on Hebrew, which is both considerably lower-resource compared to the other six test languages, and uses its own unique writing system.

5.1 Correctness feature captures possession

Figure 8 shows how choosing samples of an image generated by a model, elicited by a high- or low-**correctness** score naturally reveals in which languages which concepts are possessed (e.g., for Stable Diffusion 2, ES:rabbit, JA:snow, and ID:guitar are possessed. When a model possesses a concept, the outputted images are often visually similar with the tangible concept set in similar scenarios, and are.

5.2 Types of concept non-possession

A model *not* possessing a concept can manifest in a few different scenarios we identified in Figure 8 (b). DALL-E 2 doesn't possess "prince" in English because it outputs a variety of different images, including human portrait photos, pictures of pictures, toys, and dogs. These *non-specific* error cases are probably reflective of overall ill-defined concepts.

A second type of possession failure we observe, we dub *specific collisions*. For example, Figure 1 and Figure 3 show JA collisions for the DALL-E mini/mega family. Both models consistently generate images of humans for "dog" but pictures of green landscape scenes for "airplane." While these generated concepts are incorrect, they represent an incorrect mapping to a different concept rather than a mere lack of conceptual possession. We also observe cases where specific collisions only occur part of the time, such as in the case of DALL-E 2 and ZH:ticket (Figure 8 (b)).

Finally, we observed cases of *generic collisions*. For example, DALL-E 2 consistently generates

Model	EN		ES		DE		ZH		JA		HE		ID		Avg	
	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc	Xc	Wc
DALL-E Mega	**81**	**28**	**65**	26	**64**	**26**	29	21	32	21	28	19	**51**	**25**	50	**24**
DALL-E Mini	78	27	59	25	50	23	33	21	31	21	34	**20**	49	24	48	23
SD 1.1	69	26	52	23	46	22	32	19	37	21	28	17	39	21	43	21
SD 1.2	71	26	48	23	44	22	28	19	35	21	24	17	37	21	41	21
SD 1.4	69	26	46	23	40	22	26	20	34	21	24	17	34	21	39	21
SD 2	76	27	54	24	51	24	34	19	31	21	29	17	37	21	45	22
CogView 2	37	20	42	20	39	20	**62**	**25**	40	21	**38**	**20**	42	20	43	21
DALL-E 2	61	27	55	**27**	54	**26**	44	**25**	42	22	36	19	42	**25**	48	24
AltDiffusion m9	64	26	59	25	49	22	55	**25**	**55**	**25**	**38**	**20**	43	22	**52**	23
Avg	67	26	53	24	49	23	38	22	38	21	31	18	42	22		

Table 2: Correctness scores (Xc and Wc) averaged for all concepts within a column language for all models. Note that Xc for CogView2 is relative to ZH rather than EN. AltDiffusion performs best in terms of total average Xc, and number of Xc or Wc column "wins." DALL-E Mega performs best on Latin languages and avg Wc.

images of desert or Mediterranean scenery when prompted with "eye" in Hebrew (Figure 8 (b)). This pattern shows up across a diverse set of models and prompts. Figure 1 shows how across "dog," "airplane," and "face," DALL-E mega, Stable Diffusion 2, and DALL-E 2 seem to generate vaguely-Israel-looking outdoor scenes regardless of eliciting concept. This is probably reflective of a small sample-size bias in the training data. Hebrew in general is the prime exhibitor of generic collision cases in our study.

5.3 Model comparison

Table 2 shows the the use of correctness scores in the ***CoCo-CroLa*** benchmark to compare the 9 models. As expected, given its multilingual training regimen, AltDiffusion m9 outperforms the other T2I models on average, and in terms of total wins. It is particularly strong relative to the other models in Japanese and Chinese (with the exception of the Chinese-only CogView 2, which is best on Chinese but worst on average overall for both Xc and Wc).

However, despite the strong average performance of AltDiffusion, there's a lot of room for improvement. For example, its improvements in terms of JA and HE performance come at a cost of significantly reduced EN and DE performance relative to Stable Diffusion 2, its initialization checkpoint. The ***CoCo-CroLa*** benchmark can guide future work in adapting T2I models to further multilinguality without losing conceptual coverage on source languages.

5.4 Distinctiveness captures generic collisions

Figure 9 shows the distribution of the inverse distinctiveness score Dt. On this plot, more rightward probability mass indicates a distribution of concepts for which distinctiveness is **low** relative to a generic sample of images produced by a given model in that language. The four Latin script languages (EN, ES, DE, ID) exhibit the lowest inverse distinctiveness, and are thus the least prone to producing generic failure images. Hebrew is an outlier in terms of concept-level Dt, with a high inverse distinctiveness.

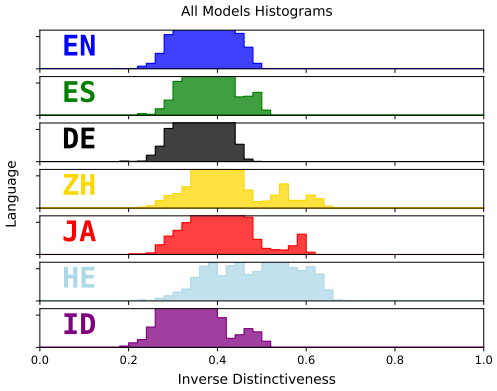
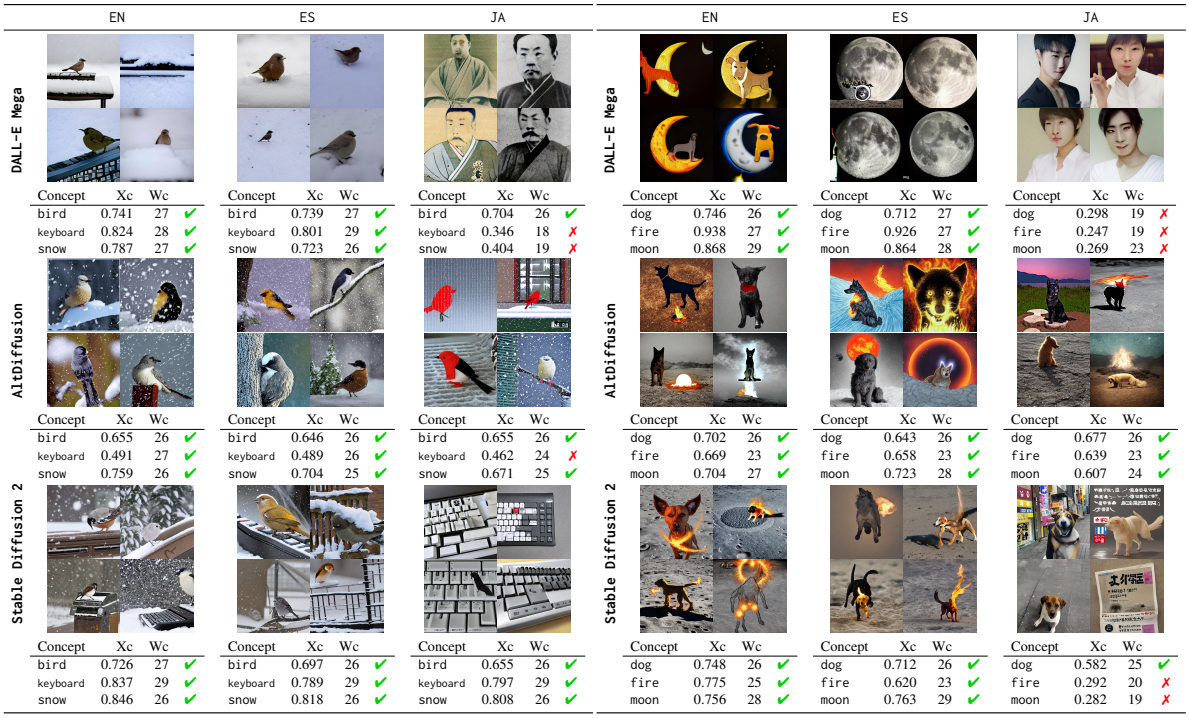


Figure 9: Histograms of the inverse distinctiveness scores for all models and all concepts.

5.5 Ranking concepts by Xc

For a specific model and language, ***CoCo-CroLa*** can be used as a concept-level analysis tool. For example, by performing the same ranking over a specific (model, language) pair, we can find the most well-covered and poorly-covered concepts for that pair. For all models and languages, an interactive ranking demo based on ascending and descending Xc and Wc is available at (anonymized) conceptualcoverage.github.io/. For example, we found "snow" to be a concept possessed in EN and ES for DALL-E Mega, AltDiffusion, and Stable Diffusion, but only possessed in JA by Stable Diffusion 2. Similarly for "dog" and "fire," but with respect to AltDiffusion.



(a) "a bird using a keyboard in the snow"

(b) "a dog made of fire standing on the moon"

Figure 10: Cross-model analysis of more complicated, creative prompts combining concepts including "snow," "keyboard," "bird," "dog," "fire," and "moon." We find that **if a model is found to not possess a concept, it will not be able to produce more complicated prompts including the concept.** This validates ***CoCo-CroLa*** as an efficient way to capture an overview of a model's generalization capabilities.

5.6 Concept possession as a proxy

In this section we will discuss how **a lack of coverage of a concept implies an inability for a model to use it in more complex, creative phrases,** validating ***CoCo-CroLa***'s paradigm.

To investigate this we manually translated two creative prompts including concepts found to be differentially present in DALL-E Mega, AltDiffusion, and Stable Diffusion subsection 5.5 from English into Spanish and Japanese. The prompts were: "a bird using a keyboard in the snow," (ES: "un pájaro usando un teclado en la nieve," JA: "雪にキーボードを使っている鳥") and "a dog made of fire standing on the moon," (ES: "un perro hecho de fuego pisando en la luna," JA: "火でできた犬が月に立っている").

Figure 10 clearly shows that, using thresholds for non-possession of Xc < 0.5 and Wc < 25, **if a concept is not possessed by a model according to *CoCo-CroLa*, it will be unable to successfully generate creative images containing it**.

However, other capabilities including compositionality and perhaps a sort of *verb-level conceptual possession* are probably required in order to make the converse (possession implies capability to generate creatively) to be true. This is a promising direction for future work. This suggests that evidence of concept-level coverage in a model can be used as a proxy for generalization capabilities to more complex prompts containing the concept, at least in the case of tangible noun concepts. This is good news, as it **enables assessment of the infinite space of creative prompts from a feasible, constrained set of concepts.**

6 Conclusion

Multilingual analysis of text-to-image models is desirable for both improving the multicultural accessibility of T2I systems and deepening our understanding of their semantic capabilities and weaknesses. Analyzing a model's *conceptual coverage* is a simple and straightforward way to do this.

We demonstrated that these concepts are core building blocks for producing impressive images, and that analyzing them is a useful proxy for assessing the truly impressive capabilities of T2I models.

Our technique, ***CoCo-CroLa*** is a first step toward further work in this domain. Utilizing our technique, larger benchmarks containing more languages and concepts can easily be built.

Limitations

The ***CoCo-CroLa*** benchmark generating procedure is intended to be a multilingual evaluation that can be scaled to even larger sets of concepts and languages without experienced annotators. In the interests of both concept and language quantity scale, we opted for an automated procedure which leverages machine translation systems, can introduce translation errors. In subsection 4.1 we describe how we leverage existing multilingual knowledge graphs to mitigate mistranslations. However, some errors still make it through the pipeline. For example, "flame" was translated into spanish as "llama" rather than "flama." Debugging and reducing errors of this kind is a direction for future work.

Additionally, typological variation between languages can introduce complications in applying our framework. For example, while simple template filling for prompting is straightforward in Chinese, which requires no word-dependent articles, in English phonological properties of the word govern the preceding article, and in Spanish and German grammatical gender do the same. Hebrew has gendered nouns, adjectives, and verbs but not articles, on the other hand. Overall, it appears that these have limited influence as grammaticality isn't a crucial feature in the prediction of image tokens performed in T2I models, Appendix B.

Ethics Statement

Images of human faces are generated by our model. To mitigate the minor risk of resemblance to real people, we have downsampled all images. However, we believe this risk is mitigated by the lack of personal names in the querying data. Furthermore, we believe demonstrating that human faces are generated and under which conditions they are is important for documentation of bias and harm risks in these models.

Our data is distributed under the Wikipedia CC license.

References

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*.

Hyundong Cho, Chinnadhurai Sankar, Christopher Lin, Kaushik Ram Sadagopan, Shahin Shayandeh, Asli Celikyilmaz, Jonathan May, and Ahmad Beirami. 2021. Checkdst: Measuring real-world generalization of dialogue state tracking performance. *arXiv preprint arXiv:2112.08321*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh. 2021. Dall·e mini.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in scene description-to-depiction tasks. *AACL*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. 2022. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

VU Prabhu and A Birhane. 2020. Large datasets: A pyrrhic win for computer vision. *arXiv preprint arXiv:2006.16923*, 3.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Michael Saxon, Xinyi Wang, Wenda Xu, and William Yang Wang. 2022. Peco: Examining single sentence label leakage in natural language inference datasets through progressive evaluation of cluster outliers. *arXiv preprint arXiv:2112.09237*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A multitask benchmark and analysis platform for natural language understanding. *Advances in Neural Information Processing Systems*, 32:3261–3275.

Wenda Xu, Yilin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not all errors are equal: Learning text generation metrics using stratified error synthesis. *arXiv preprint arXiv:2210.05035*.

A Mitigating translation errors

Source language term lists. We first produce a list of English nouns by collating words in term frequency lists extracted from TV closed captions and contemporary fiction novels from Wiktionary[4], and filter for the 2000 most frequent words in this combined list, and augment it with class label names from CIFAR100 (Krizhevsky et al., 2009).

Finding good translations. We feed the list English words into a custom translation pipeline, which simultaneously queries BabelNet (Navigli and Ponzetto, 2010), and an ensemble of four commercial translation systems: Google Translate, Bing Translate, Baidu Translate, and iTranslate[5].

In response to an English query, the BabelNet API returns a collection of "SynSets," subgraphs of a combined multilingual word and entity graphs centered on a node the query word maps to (see Figure 5 for examples). Each subgraph links to multiple other nodes, containing terms in both the source language and the target language. These edges can represent, for example, the titles of Wikipedia articles in different language editions of Wikipedia that are marked as being equivalent, thus ensuring that by checking against SynSet edges, a degree of human validation is included automatically. The synset also contains information about whether a given word is a noun. If it is not a noun, the candidate concept is discarded.

To choose the best translation from those edges, the returned translations into the target languages of the English term from the commercial translation services are *melded* by first sorting all returns by number of languages in the return query (in the case that one translation service covers more languages than others), and filling in missing translations by prioritizing alignment in the shared language translations. If any target language is missing a word for a concept at the conclusion of this process, that concept is discarded from the final list.

Post-filtering. Once a list of melded translations from the commercial service is returned, each row in the candidate aligned concept list is checked against the corresponding BabelNet SynSets to ensure each translation is present as a connected node, for pseudo-human evaluation. At the end of this process, a list of approximately 250 concepts is returned. Finally, we manually remove terms that are verb-noun collisions (e.g. hike) to ensure this ambiguity didn't drive any poor translations. The final list for ***CoCo-CroLa*** v0.1 contains 193 concepts.

B Validating the prompt templates

As mentioned in subsection 4.2, the simple template-based approach to generating prompts for concepts leads to the introduction of grammatical errors, e.g. "a photograph of dog."

However, it is questionable whether small grammatical or logical errors like missing articles matters for high-resourced, well-covered languages like English. After all, the models are clearly able to generate high quality "photograph of dog" pictures without the word "a" in the sentence (Figure 1). But, does the prompt phrasing matter for lower-performance languages in a model? To investigate the impact of prompt phrasing on conceptual coverage, we tested a variety of English, Spanish and Chinese prompt phrasings on the concepts "dog," "sea," "airplane," and "ship" (selected for their wide distribution across the cross-correlation correctness metric range).

For the English prompts, we experimented with

[4] `en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Contemporary_fiction`, `.../TV/2006/1-1000`

[5] Using the `translators` PyPi package.

including the articles “a,” “the,” “my,” and “an,” as well as using the words “photograph,” “image,” “photo,” and “picture.” For Spanish, we used variations on the phrase “un foto de” (a photo of), including the same set of articles in English “un/una,” “el/la,” “mi,” “tu,” (your) and “nuestra/o” (our). For Chinese, we tried examples that both included and excluded the possessive particle “的” (de), as well as the words “照片” (zhaopian) and “图片” (tupian) for picture/photograph, and including or excluding the prepended phrase “一张” (yi zhang) to create the meaning “one photograph.” We reran the full 193 concept image generations in those three languages for Stable Diffusion 2 and AltDiffusion.

We found limited impact across all of these dimensions. Full details available in our anonymous demo at (anonymized) conceptualcoverage.github.io/.