Lost in Translation? Translation Errors and Challenges for Fair Assessment of Text-to-Image Models on Multilingual Concepts

Anonymous ACL submission

Abstract

Benchmarks of the multilingual capabilities of text-to-image (T2I) models compare generated images conditioned on test language and then compare the results with the expected image distribution. One such benchmark, "Conceptual Coverage Across Languages" (CoCo-CroLa), assesses the tangible noun inventory of T2I models by prompting them to generate pictures of them in seven input languages and comparing the output image populations. Unfortunately, we find that this benchmark contains translation errors of varying severity in Spanish, Japanese, and Chinese. We provide corrections for these errors and analyze how impactful they are on the utility and validity of CoCo-CroLa as a benchmark. We reassess multiple baseline T2I models with the revisions, compare the outputs elicited under the new translations to those conditioned on the old, and show that a correction's impactfulness on the image-domain benchmark results can be predicted in the text-domain using similarity metrics. Our findings will guide the future development of T2I multilinguality metrics by providing analytical tools for making practical translation decisions.

1 Introduction

With growth in the popularity of generative text-to-image (T2I) models has come interest in assessing their capabilities across many dimensions, including multilingual accessibility. The CoCo-CroLa (Saxon and Wang, 2023) benchmark attempts to capture how well "concept-level knowledge" within a T2I model is accessible across different input languages. It compares the output image populations of a system under test when prompted to generate images of a tangible concept in a *test language* to the images generated from a semantically equivalent prompt in a *source language*. It and similar benchmarks rely on correct translations for validity, lest "possessed" concepts be mistakenly assigned false negatives.

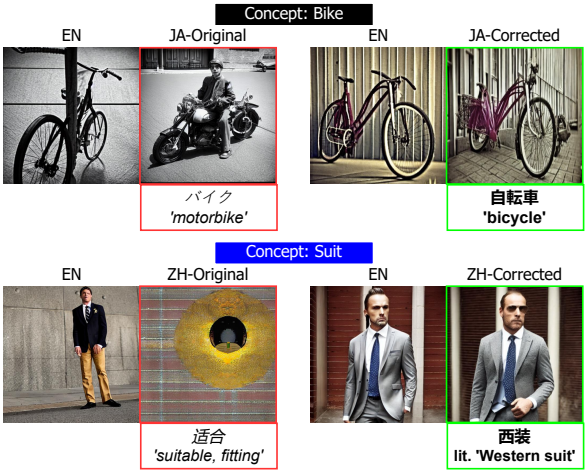


Figure 1: The CoCo-CroLa benchmark mistranslated concepts such as *bike* in JA and *suit* in ZH. With the correct translations (right) AltDiffusion does in fact "possess" them; originally (left) they were false negatives.

We find a strict *error candidate rate* of 4.7% for Spanish, 8.8% for Chinese, and 12.9% for Japanese in the CoCo-CroLa v1 (CCCL) concept translations through manual analysis by fluent speakers. **These error candidates are not filtered by severity**. While some candidates are severe translation errors that drive false negatives (Figure 1), others are marginal annotator disagreements that might not matter (Table 1). In this work, **we study when and why mistranslations actually impact CCCL results** to improve future T2I multilinguality benchmarks. We:

1. Provide corrections for CCCL in ES, JA, and ZH, and evaluate four T2I models with them
2. Introduce a text-domain mistranslation severity metric ΔSEM that is predictive of the impact of a mistranslation correction on the improvement of model performance
3. Analyze the future work in machine translation for the T2I model assessment domain

Error Type	Concept	Lang.	Original	Corrected	Reason for Correction
Transliteration	Rock	JA	ロック	岩	ロック, *rokku*, refers principally to "rock music" instead of stones in nature.
	Flame	ES	*llama*	*flama*	*Llama*, though a correct translation for "flame," coincides with the animal in English.
Wrong Sense	Ground	JA	接地	地面	接地 refers to the concept of grounding in electronics.
	Table	ZH	表	桌子	表 means a tabular form or a spreadsheet, not a four-legged furniture.
Ambiguity	Milk	JA	乳	牛乳	乳 may mean breast or any kind of milk. 牛乳 means the milk produced by cows.
	Tent	ES	*tienda*	*...de acampar*	*Tienda* alone more often means "store," *tienda de acampar* specifies (camping) tent.
Formality	Teacher	JA	先生	教師	先生 is a common title to address an educated person, e.g., teacher, doctor, lawyer.
	Father	ZH	爸爸	父亲	爸爸 is the colloquial addressing equivalent to 'daddy'. 父亲 is more formal.

Table 1: Examples of the translation errors found in the original CoCo-CroLa benchmark in Japanese (JA), Chinese (ZH), and Spanish (ES). See Appendix A.1 for our definitions of each error type along with all errors.

2 Motivation & Approach

The **CoCo-CroLa** benchmark (CCCL) evaluates a T2I model's ability to generate images of an inventory of tangible concepts when prompted in different languages (Saxon and Wang, 2023). Given a tangible concept c, written in language ℓ as phrase c_ℓ, the i-th image produced by a multilingual T2I model f on the concept c_ℓ can be expressed as:

$$I_{c_\ell,i} \sim f(c_\ell) \quad (1)$$

The images generated in language ℓ are considered *correct* if they are faithful to their equivalent counterparts in the source language ℓ_s. The **CCCL Score** a.k.a the **Correctness Metric** regarding a single concept c is conveyed as the cross-consistency $X_c(f, c_\ell, c_{\ell_s})$:

$$X_c = \frac{1}{n^2}\sum_{i=0}^{n}\sum_{j=0}^{n}\text{SIM}_F(I_{c_\ell,i}, I_{c_{\ell_s},j}) \quad (2)$$

where we sample n images per-concept per-language (we use 9), and $\text{SIM}_F(\cdot,\cdot)$ measures the cosine similarity in feature space by image feature extractor F. In practice, the default source language ℓ_s is English and F is the CLIP visual feature extractor (Radford et al., 2021).

2.1 Translation Errors in CoCo-CroLa

CCCL requires correct translations of each concept c from the source language ℓ_s into a set of semantically-equivalent translations in each test language ℓ. Saxon and Wang (2023) built CCCL v1's concept translation list using an automated approach so as to allow "new languages to be easily added" without experts in each new language.

They use an ensemble of commercial machine translation systems to generate candidate translations and the BabelNet knowledge graph (Navigli and Ponzetto, 2010) to enforce word sense agreement. Unfortunately, this approach introduces translation errors (Table 1).

We check the Spanish, Chinese, and Japanese translations using a group of proficient speakers, following a protocol described in Appendix A.4, who identify a set of *translation error candidates* that may not sufficiently capture a concept's intended semantics in English, for various reasons.

Some of the candidate errors, such as the error for *rock* in JA (Table 1), represent severe failures to translate a concept into its common, tangible sense—it is incoherent to test a model's ability to generate pictures of rocks by prompting it with "rock music." However, other candidate errors, such as *father* in ZH are still potentially acceptable translations, but deviate from the annotators' preferred level of formality or specificity.

To decide which corrections ought to be integrated in future T2I multilinguality benchmarks, quantifying both the significance of each translation correction is and its impact on the CCCL score for its concept is desirable.

2.2 Quantifying Error Correction & Impact

Characterizing the *impact* of a translation correction on model behavior is simple; we check ΔX_c, the change in the CCCL score going from the original concept translation c_ℓ to the corrected c'_ℓ,

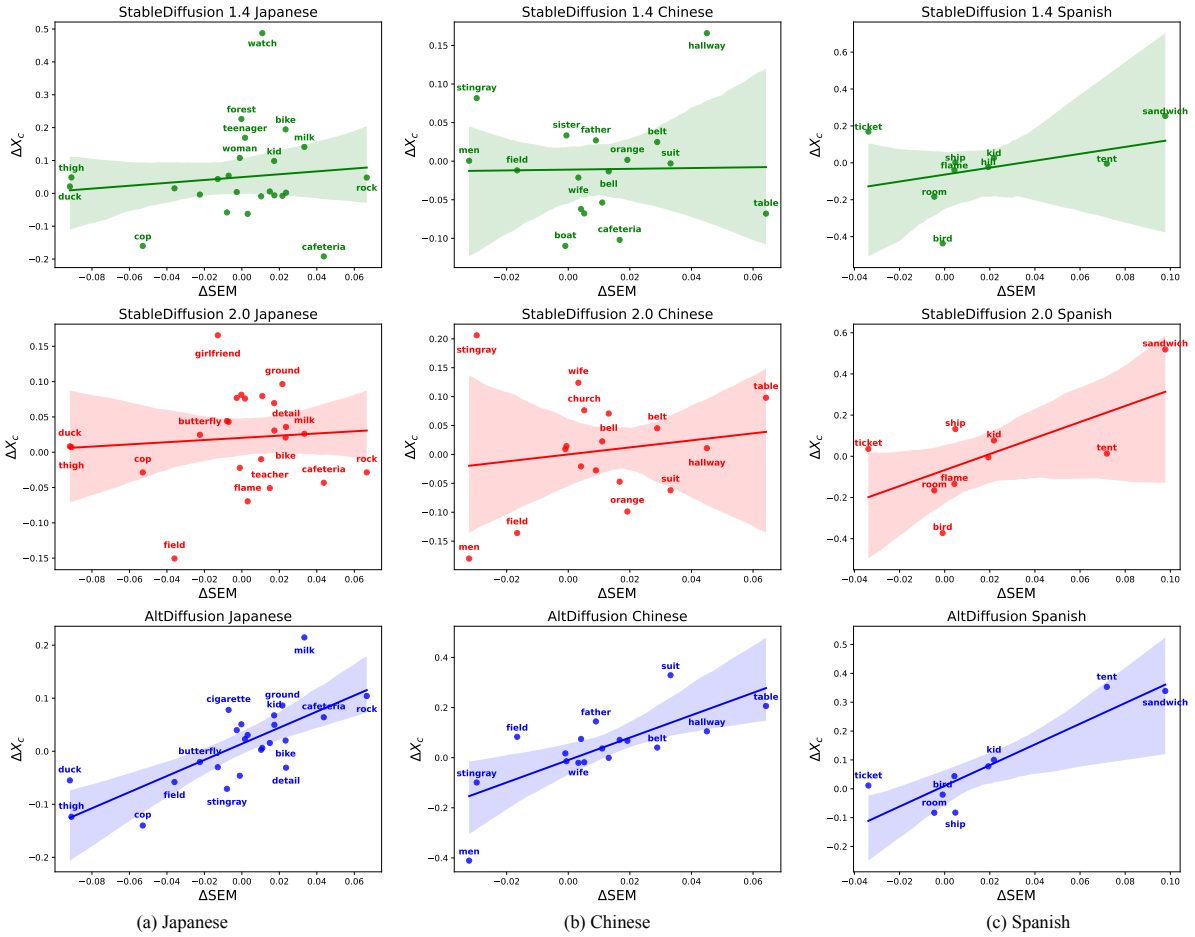
$$\Delta X_c(c,\ell) = X_c(f, c'_\ell, c_{\ell_s}) - X_c(f, c_\ell, c_{\ell_s}) \quad (3)$$

by comparing the generated population of images elicited from the corrected term $I_{c'_\ell}$ to the candidate translation error-conditioned images I_{c_ℓ}.

We quantify the significance of the translation correction as the *improvement in semantic similarity* $\Delta\text{SEM}(c_{\ell_s}, c_\ell, c'_\ell)$ using text feature extractor F_t and cosine similarity metric $\text{SIM}(\cdot,\cdot)$

$$\Delta\text{SEM} = \text{SIM}_{Ft}(c_{\ell_s}, c'_\ell) - \text{SIM}_{Ft}(c_{\ell_s}, c_\ell) \quad (4)$$

We use embeddings from the multilingual SentenceBERT (Reimers and Gurevych, 2019) text embedder OpenAI CLIP-ViT-B32 model as F_t.



(a) Japanese (b) Chinese (c) Spanish

Figure 2: Scatterplots showing the impact of the corrections to each concept in JA, ZH, and ES on the conceptwise improvement to the CCCL cross-consistency score, ΔX_c, as a function of ΔSEM.

3 Results & Analysis

We generate output images using StableDiffusion 1.4, 2.0, 2.1[1] (Rombach et al., 2022) and AltDiffusion (Chen et al., 2022), for all concepts corrected by our annotators, in English, Spanish, Chinese, and Japanese, using both the original concept translations c_ℓ from CoCo-CroLa v1 (Saxon and Wang, 2023) and the corrected translations c'_ℓ. Model details are provided in Appendix A.3.

Figure 2 shows the relationship between ΔSEM and ΔX_c for all corrected concepts for StableDiffusion 1.4 and 2.0, and AltDiffusion.

It would be reasonable to suppose that *corrections are most useful in languages that a model actually "knows."* After all, correctly-translated Klingon should be just as incomprehensible to a non-Klingon model as incorrect Klingon. Our Figure 2 correlation findings support this hypothesis.

Note the pronounced, significant positive correlation between the two variables for AltDiffusion in all languages (third row of Figure 2) and in Spanish for all models (third column). These model/language pairs (JA/AltDiffusion, ES/StableDiffusion 2.0, etc) were all found by Saxon and Wang (2023) to be "well-possessed" (high average Xc across the mostly correct concepts) in CoCo-CroLa v1.

StableDiffusion 1.4 was trained on the primarily-Latin script LAION-en-2b (Schuhmann et al., 2021), and thus lacks capabilities in non-Latin script languages JA, ZH. Consequently, there is no significant relationship between more semantically divergent corrections with high ΔSEM and larger improvements to concept correctness ΔX_c for SD 1.4 on those languages. Meanwhile, AltDiffusion, which conditions output images on a multilingual encoder XLM-Roberta (Conneau et al., 2020), benefits from corrections in all languages with a significant correlation.

Unfortunately our understanding of the connection between corrections and performance improvements is limited by few available corrections.

[1] Plots for StableDiffusion 2.1 in Appendix Figure 5.

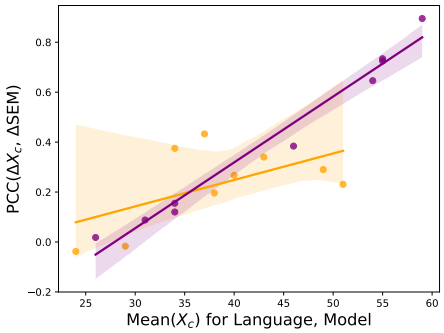


Figure 3: For a (model, language) pair we observe a positive correlation between the Pearson's correlation coefficient of ΔSEM and ΔX_c, and the average X_c (proxy for the model's language ability). This holds for both **real corrections** and **pseudocorrections**.

We sidestep this issue with a *pseudocorrection experiment*, where we generate synthetic erroneous *pseudo-original translations* by sampling 10 other concepts, which we "correct" with the original CCCL translations for German, Indonesian, and Hebrew. For example, we assign the concept *eye* the Indonesian word *guru* (EN:teacher) as the pseudo-original. We then "correct" this word to *mata*, the original correct translation, and assess ΔXc and ΔSIM with c_{ℓ_s}:*eye*, c_ℓ:*guru* and c'_ℓ:*mata*.

This gives us 1,930 ΔXc, ΔSIM pairs for each language and model (plot in Appendix Figure 6). We report the PCC for each of these pairs along with the average CCCL Xc reported in Saxon and Wang (2023) in Figure 3. We indeed find that the same relationship for the real corrections holds for pseudocorrections. **Thus, text-only multilingual semantic similarity features are predictive of the measurable importance of a translation correction on the output image distribution.**

4 Discussion & Conclusions

Subjectivity. A reliable T2I multilinguality assessment must report true possession failures—examples where a model fails to generate correct images of a concept, when it is correctly prompted to do so. Correct translations are required.

Unfortunately, the problem of choosing one "correct translation" necessarily contains subjectivity. This study was an attempt to tackle this subjectivity by casting a wide net of error candidates, and taking the corrections that proved impactful.

The consequential *benchmark errors* that we found were mainly *false negatives* where a mistranslation caused a concept to be erroneous marked as not-possessed (Figure 1).

Image-Image Metric Blind Spots. We observed interesting borderline (potential false positive) cases where CoCo-CroLa scored mistranslated concepts as possessed. For example, *bike* in Japanese. Figure 1 shows that under the erroneous translation, AltDiffusion generates pictures of *motorcycles* rather than *bicycles* as it does in English. However, **Xc doesn't actually change much under this correction** as shown in Figure 2, Table 3. **The CLIP similarity score employed by CCCL is functionally blind to the difference** between a bicycle and motorcycle. In a way, the metric is robust to its own mistranslation because the img-img similarity metric attends to structural similarities between the specificity-misaligned meanings.

Tangible object translation as an MT domain. Single word concepts are out of distribution for how machine translation models are typically trained. By providing the individual English tangible nouns as input Saxon and Wang (2023) were expecting an unreasonable amount of implicit commonsense reasoning from commercial MT systems—the correct sense out of many had to be selected for success. Furthermore, their use of the BabelNet knowledge graph as a consensus mechanism also can reinforce sense errors. For example, the *rock* sense error for JA (music genre rather than physical object, Table 3, Figure 4) was also present in Hebrew, probably due to shared edges (SynSets) in the knowledge graph.

Solving Mistranslations. Future benchmarks should leverage contextualized sentences as input to the MT models (eg, "watch for falling rocks") rather than the decontextualized word alone to improve robustness. LLMs should be employed rather than knowledge graphs for merging.

5 Related Work

Prior work such as Drawbench (Saharia et al., 2022), DALL-Eval (Cho et al., 2022), and T2I-CompBench (Huang et al., 2023) all evaluate the capabilities of T2I models. Prior works on errors in vision-language benchmarks include Agrawal et al. (2018) finding spurious correlations in the training data of VQA (Antol et al., 2015), Luo et al. (2022) filtering out unsolvable cases in Who's Waldo (Cui et al., 2021), and Ye and Kovashka (2021) exploiting repeated texts in questions and answers to achieve high performance in Visual Commonsense Reasoning (Zellers et al., 2019).

Limitations

Trivially, human annotators for every language would remove false-negative mistranslations from future benchmarks, but this has drawbacks. There is a trade-off between scalable broad-net representation (and the identification of *potential* collisions and other offensive errors identified in Saxon and Wang (2023)) and certainty of correctness.

Our work incorporates human efforts of proficient foreign language users to correct the translation errors caused by the machine translation pipeline in the original CoCo-CroLa benchmark. This could potentially bring human biases into the nuance of factors such as words' choices, introducing less culturally neural expressions as a result.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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*, abs/2211.06679.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL 2020*, pages 8440–8451, Online. Association for Computational Linguistics.

Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. 2021. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1374–1384.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*.

Yiran Luo, Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2022. To find waldo you need contextual cues: Debiasing who's waldo. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, page 355–361, Dublin, Ireland. Association for Computational Linguistics.

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, Uppsala, Sweden. Association for Computational Linguistics.

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 *ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 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 *CVPR 2022*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. 35:36479–36494.

Michael Saxon and William Yang Wang. 2023. Multilingual conceptual coverage in text-to-image models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4831–4848, Toronto, Canada. Association for Computational Linguistics.

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*, abs/2111.02114.

Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3181–3189.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

A Appendix

A.1 Definition of error types with examples

Transliteration. The translated term is a direct transliteration of the source concept in pronunciation, but it carries a different meaning. For example, the transliteration of *Rock* in Japanese is commonly related to 'Rock Music', rather than stones found in nature.
Wrong Sense. The translated term picks an alternative (and often less tangible) sense from the source concept. For example, the original Chinese translation for *Table* diverges to the sense of 'spreadsheet, tabular', instead of the presumptive home furniture item.
Ambiguity. The translated term introduces a word with multiple meanings from the unambiguous source concept. For example, the Japanese translation for *Milk* originally uses a single character that can mean any kind of animal or human milk, or even the organ of the breast.
Formality. The translated term uses an expression in an improper formality. For example, the original Chinese translation for *Father* is only heard in casual conversations.

A.2 Additional Resource Information

License and Terms We follow the same license and terms of the original CoCo-CroLa benchmark.

Intended Use Our dataset is intended to evaluate the performance of text-to-image generation models.

Offensive Content Some of the translations we found can lead to offensive images, e.g. the original translation for "Milk" in Japanese can also mean breast.

A.3 Computational Experiments Details

Dataset Statistics We provide a collection of 193 multilingual concepts in 6 languages. We have also modified 50 of them with verified translations by human annotators.

Models Employed See Table 2.

Model	Param. Count	Repository	Training Language
StableDiffusion 1.4	860M	HF:CompVis/stable-diffusion-v1-4	No language filter (en)
StableDiffusion 2	NA	HF:stabilityai/stable-diffusion-2	No language filter (en)
StableDiffusion 2.1	NA	HF:stabilityai/stable-diffusion-2	No language filter (en)
AltDiffusion m9	1.7B	HF:BAAI/AltDiffusion-m9	EN, ES, FR, IT, RU, ZH, JA, KO

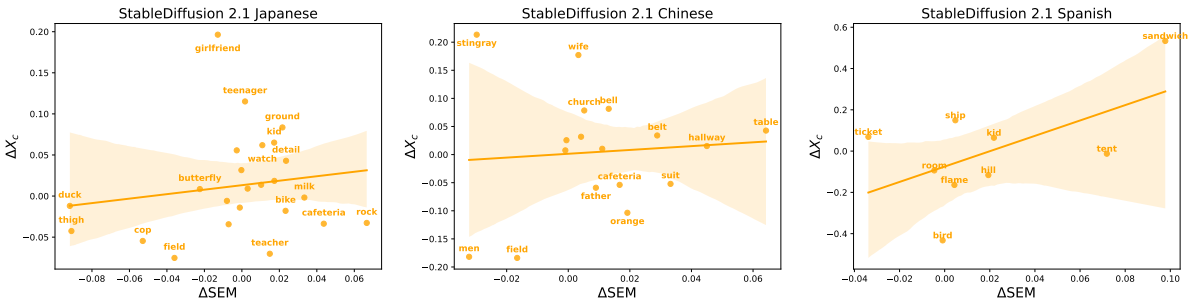
Table 2: The set of text-to-image models we evaluated with (Table adapted from (Saxon and Wang, 2023).

Experimental Setup No hyperparameter search was necessary as we did not train a model. We generated 9 images for each (language, model, concept) triple.

A.4 Human Annotator Details

We asked graduate students who were speakers of the languages to check each concept in CoCo-CroLa. The annotators were volunteers who spent about 10 minutes on the annotation task. We then checked the annotations against bilingual English-{Spanish, Japanese, Chinese} dictionaries.

Figure 4: Qualitative examples of selected mistranslated concepts found in Coco-CroLa generated by AltDiffusion and multiple versions of Stable Diffusion - **Top left:** "Rock" in Japanese, **Top right:** "Suit" in Chinese, **Bottom left:** "Tent" in Spanish, **Bottom right:** "Table" in Chinese. Noticeably, we observe that T2I models such as Stable Diffusion 2 do not benefit from correcting the translations, as their outputs in the aforementioned languages remain irrelevant similarly to using random prompts.



(a) Japanese (b) Chinese (c) Spanish

Figure 5: Scatterplots showing the impact of the corrections to each concept in JA, ZH, and ES on the conceptwise improvement to the CCCL cross-consistency score, ΔX_c, as a function of ΔSEM, for StableDiffusion 2.1 .

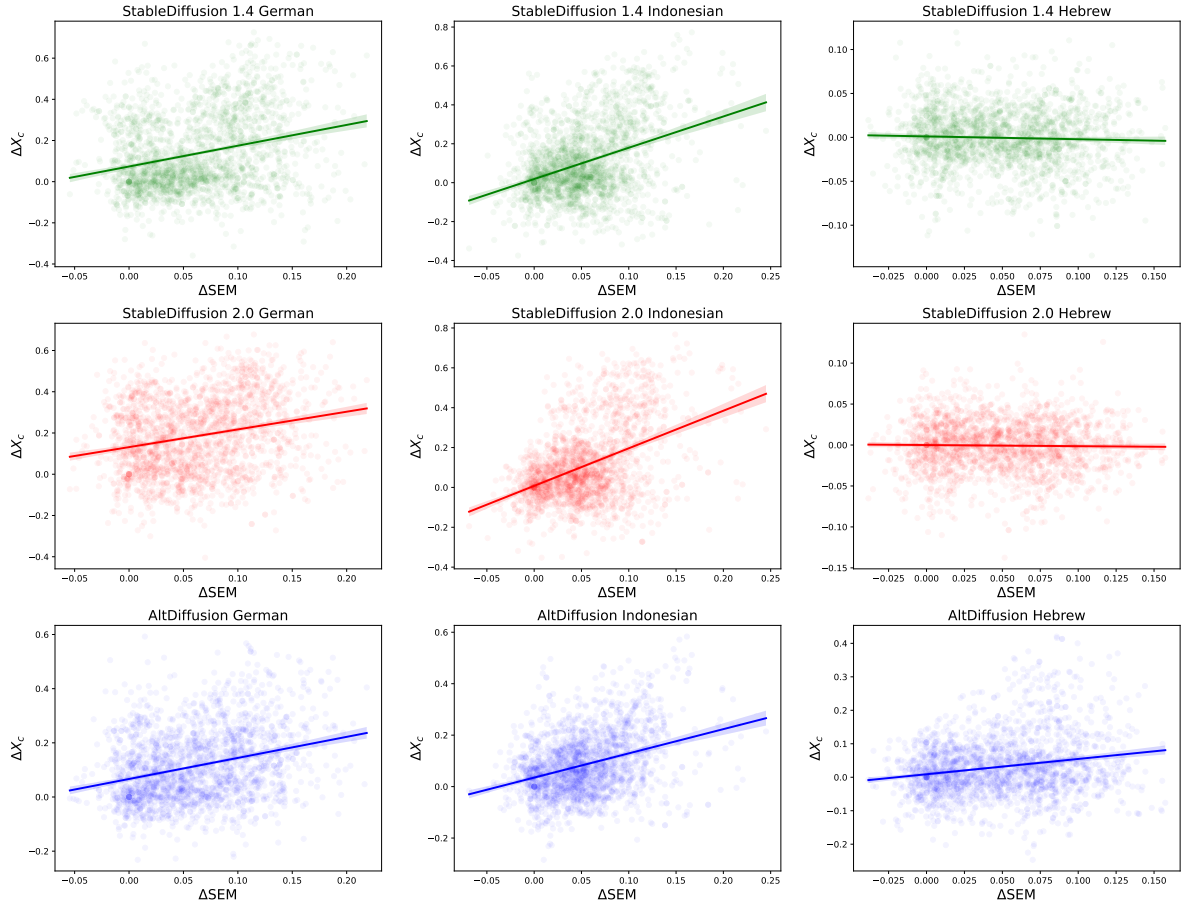


Figure 6: Scatterplots for the pseudocorrection experiments. Transparent circles are used to make distribution mass more visible.

Concept	Original	Corrected	ΔSEM	ΔX_c (CCCL Improvement) for model SD 1.4	SD 2	SD 2.1	AD
Below are the mistranslated concepts in Japanese.							
duck	鴨	アヒル	-0.092	0.021	0.008	-0.012	-0.055
thigh	腿	ふともも	-0.091	0.048	0.007	-0.043	-0.124
cop	警官	お巡りさん	-0.053	-0.160	-0.029	-0.055	-0.140
field	分野	田んぼ	-0.036	0.015	-0.151	-0.075	-0.058
butterfly	蝶	蝶々	-0.022	-0.004	0.025	0.009	-0.020
girlfriend	ガールフレンド	彼女	-0.013	0.044	0.166	0.196	-0.030
stingray	アカエイ	エイ	-0.008	-0.058	0.044	-0.006	-0.071
cigarette	煙草	たばこ	-0.007	0.054	0.043	-0.034	0.078
tail	尾	尻尾	-0.003	0.004	0.077	0.056	0.040
woman	女性	女	-0.001	0.108	-0.022	-0.014	-0.046
forest	森林	森	-0.000	0.226	0.081	0.032	0.051
teenager	ティーンエイジャー	少年	0.002	0.169	0.076	0.115	0.023
flame	火炎	炎	0.003	-0.062	-0.070	0.009	0.031
father	父	父親	0.010	-0.009	-0.010	0.014	0.003
watch	時計	腕時計	0.011	0.487	0.080	0.062	0.006
teacher	先生	教師	0.015	0.006	-0.051	-0.070	0.016
kid	キッド	子ども	0.017	0.098	0.070	0.065	0.068
doctor	先生	医者	0.017	-0.006	0.031	0.018	0.050
ground	接地	地面	0.022	-0.008	0.097	0.084	0.086
bike	バイク	自転車	0.023	0.195	0.021	-0.018	0.020
detail	ディテール	詳細	0.024	0.002	0.036	0.043	-0.031
milk	乳	牛乳	0.033	0.141	0.026	-0.002	0.215
cafeteria	カフェテリア	食堂	0.044	-0.192	-0.043	-0.034	0.064
rock	ロック	岩	0.067	0.048	-0.029	-0.033	0.104
Below are the mistranslated concepts in Chinese.							
men	男人	很多人	-0.032	0.001	-0.180	-0.182	-0.411
stingray	黄貂鱼	鳐鱼	-0.030	0.082	0.206	0.213	-0.099
field	领域	田野	-0.017	-0.012	-0.136	-0.184	0.083
boat	船	小船	-0.001	-0.110	0.009	0.008	0.017
sister	姐姐	姐妹	-0.001	0.033	0.014	0.026	-0.014
wife	老婆	妻子	0.003	-0.021	0.124	0.177	-0.021
bottle	瓶	瓶子	0.004	-0.062	-0.021	0.032	0.075
church	教会	教堂	0.005	-0.068	0.076	0.078	-0.018
father	爸爸	父亲	0.009	0.027	-0.028	-0.059	0.145
mouth	口	嘴	0.011	-0.054	0.023	0.010	0.037
bell	钟	铃	0.013	-0.013	0.071	0.081	-0.001
cafeteria	自助餐厅	食堂	0.017	-0.102	-0.047	-0.054	0.071
orange	橙色	橙子	0.019	0.002	-0.099	-0.104	0.067
belt	带	皮带	0.029	0.025	0.045	0.034	0.040
suit	适合	西装	0.033	-0.003	-0.062	-0.052	0.329
hallway	门厅	走廊	0.045	0.166	0.011	0.015	0.105
table	表	桌子	0.064	-0.068	0.098	0.043	0.206
Below are the mistranslated concepts in Spanish.							
ticket	boleto	billete	-0.034	0.169	0.036	0.069	0.011
room	habitación	cuarto	-0.005	-0.184	-0.166	-0.094	-0.083
bird	pájaro	ave	-0.001	-0.437	-0.373	-0.433	-0.020
flame	llama	flama	0.004	-0.040	-0.134	-0.164	0.044
ship	navío	barco	0.005	0.002	0.132	0.149	-0.083
hill	cerro	colina	0.019	-0.023	-0.005	-0.116	0.078
kid	cabrito	joven	0.022	0.027	0.077	0.065	0.100
tent	tienda	tienda de acampar	0.072	-0.005	0.013	-0.013	0.353
sandwich	emparedado	sándwich	0.098	0.254	0.519	0.534	0.339

Table 3: All identified concept translation error candidates in the original CoCo-CroLa and their corresponding corrections in Japanese, Chinese, and Spanish. Each section is sorted in ascending order of ΔSEM.