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

Michael Saxon **Yiran Luo** **Sharon Levy**
Chitta Baral **Yezhou Yang** **William Yang Wang**
University of California, Santa Barbara Arizona State University Johns Hopkins University
Equal contribution & corresponding: saxon@ucsb.edu, yluo97@asu.edu

Abstract

Benchmarks of the multilingual capabilities of text-to-image (T2I) models compare generated images prompted in a test language to an expected image distribution over a concept set. One such benchmark, "Conceptual Coverage Across Languages" (CoCo-CroLa), assesses the tangible noun inventory of T2I models by prompting them to generate pictures from a concept list translated to seven 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 with similarity scores. Our findings will guide the future development of T2I multilinguality metrics by providing analytical tools for 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 193 tangible concepts in 7 *test languages* 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.

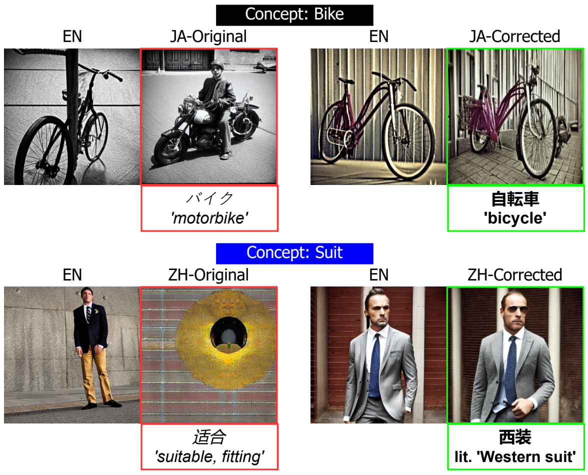


Figure 1: The CoCo-CroLa benchmark mistranslated concepts such as *bike* in JA and *suit* in ZH. With correct translations (right) AltDiffusion does in fact "possess" them; originally (left) they were false negatives.

We find a strict *error candidate rate* of 4.7% for Spanish (ES), 8.8% for Chinese (ZH), and 12.9% for Japanese (JA) 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 investigate when and why translation changes actually impact CCCL results** to improve future T2I multilinguality benchmarks. We:

1. Write *candidate corrections* for CCCL in ES, JA, and ZH, evaluated on four T2I models.

2. Introduce a text-domain comparison metric ΔSEM to predict correction significance.

3. Analyze our candidates by ΔSEM and image correctness improvement and apply impactful ones to CCCL as v1.1.

4. Report insights and considerations for future semantic T2I evaluations we uncovered.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 572–582

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

Table 1: Example error candidates from the CoCo-CroLa benchmark in Japanese, Chinese, and Spanish.

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. This is measured by the CCCL benchmark by a **correctness metric** for a single concept c as the *cross-consistency* score $X_c(f, c_\ell, c_{\ell_s})$:

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

where we sample n images per-concept per-language (we use 9), and $\mathrm{SIM}_F(\cdot,\cdot)$ 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 used 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.1.1, 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,

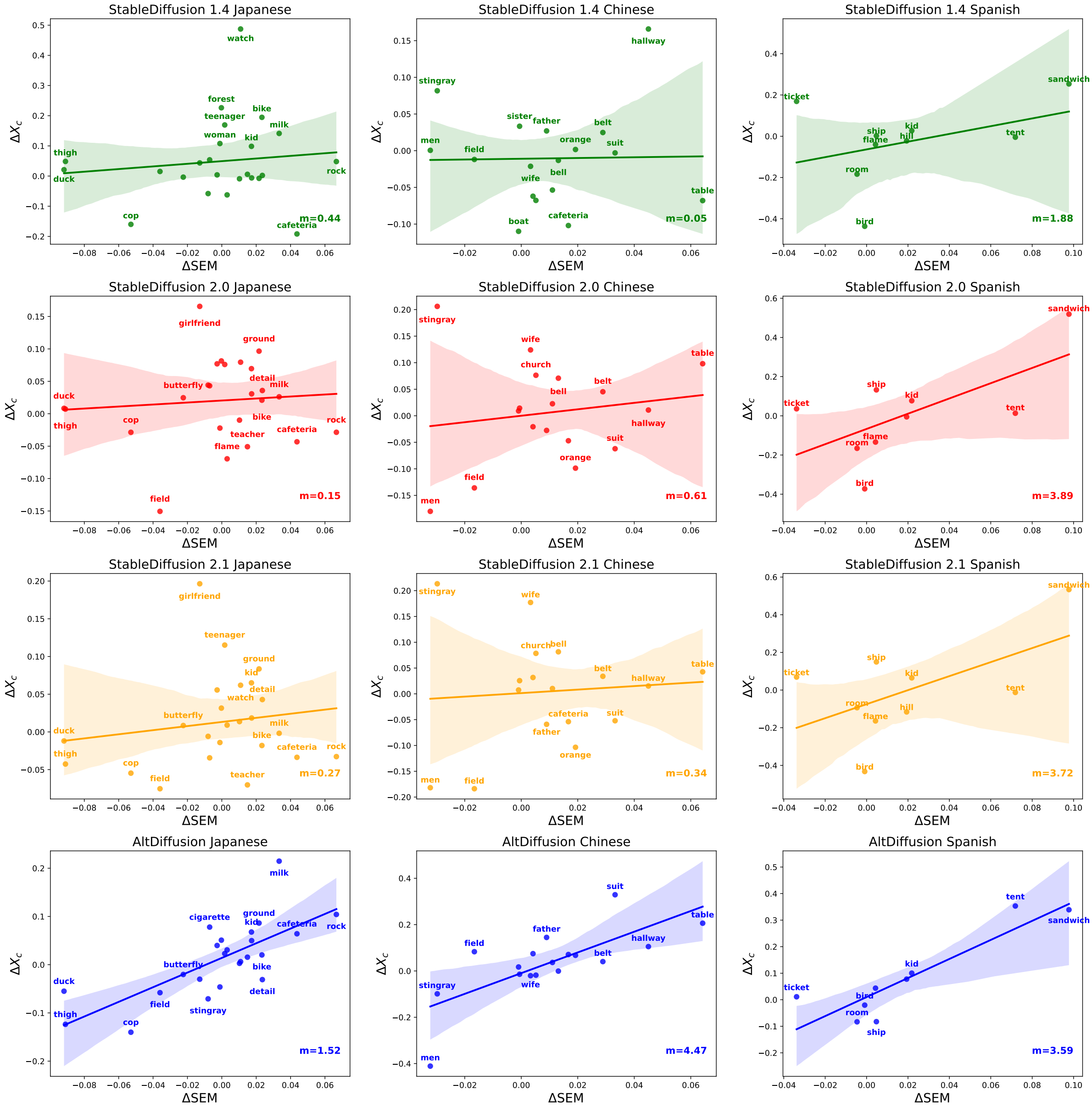
$$\Delta X_c(c,\ell) = X_c(f, c'_\ell, c_{\ell_s}) - X_c(f, c_\ell, c_{\ell_s}) \quad (3)$$

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* $\Delta\mathrm{SEM}(c_{\ell_s}, c_\ell, c'_\ell)$ using a text feature extractor F_t and cosine similarity metric $\mathrm{SIM}(\cdot,\cdot)$

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

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 correctness score, ΔX_c, as a function of ΔSEM. Slopes m at bottom-right in **bold**.

3 Results & Analyses

We generate output images using StableDiffusion 1.4, 2.0, 2.1 (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.4.

Figure 2 shows the relationship between ΔSEM and ΔX_c for all corrected concepts for StableDiffusion 1.4, 2.0, 2.1, and AltDiffusion[1]. Note the pronounced, significant positive slope of the correlations between the two variables for AltDiffusion in all languages (4th row) and in Spanish for all models (third column). Here a positive slope means that *higher-improvement* translation corrections (assessed by increased proximity to the English word in a shared embedding space) reliably correct the generated images more than the modest candidates.

These same high-slope model/language pairs (eg., JA & AltDiffusion) were found by Saxon and Wang (2023) to be "well-possessed" (high average X_c across correct concepts) in CoCo-CroLa v1. In other words, *valid corrections only matter for languages a model already "knows."* Correct Klingon is just as useless as incorrect Klingon to a non-Klingon model.

[1]Error margins are 95% regression-fit confidence intervals.

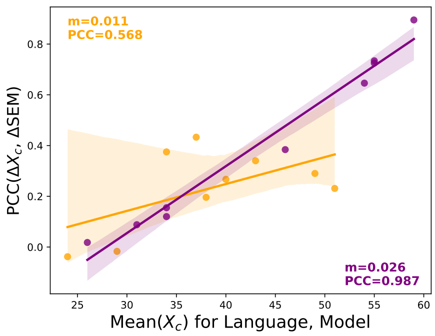


Figure 3: Languages with a high correlation between textual correction significance and image improvement (PCC) are more "well-understood" by the model (X_c), for both **real**- and **pseudo**-corrections.

Table 3 (subsection A.5) shows the same slopes m with PCCs, p-values, and intercepts for the each model and language's ΔSEM to ΔX_c relationship. The high-slope language/model pairs also tend to have higher PCC with more statistical significance.

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 and 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 the multilingual XLM-Roberta encoder (Conneau et al., 2020)–benefits from all significant corrections in all languages with statistical significance.

3.1 Pseudocorrection Experiment

Unfortunately our ability to use the aforementioned corrections to confirm our hypothesis that *T2I model language capability can be estimated from the impact of translation corrections on image-domain performance* is hindered by the small quantity of correction candidates we found. We bypass this problem with a *pseudocorrection experiment*—simulating a larger set of corrections by generating artificial errors in the other CCCL languages. We generate 10 synthetic erroneous *pseudo-original translations* for each concept in German, Indonesian, and Hebrew by randomly sampling the translations for other concepts within-language. Each concept's "correction" is its original translation.

For example, we assign the concept *eye* the Indonesian word *guru* (EN:teacher) as its pseudo-original. We then "correct" this word to *mata*, the original correct translation, and assess ΔX_c and ΔSEM with c_{ℓ_s}:*eye*, c_ℓ:*guru* and c'_ℓ:*mata*.

This gives us 1,930 ΔX_c, ΔSEM pairs for each language and model, with which we evaluate the same correlation relationship as before (plot in Appendix Figure 6). We report Pearson's correlation coefficient (PCC) for each of these pairs along with the average CCCL X_c reported in Saxon and Wang (2023) in Figure 3. The same relationship for real corrections holds for pseudocorrections, demonstrating that text-only multilingual semantic similarity features can predict the impact of a translation correction on the output image correctness.

4 Discussion & Conclusions

Our findings motivate important considerations for building future T2I semantic evaluations (Saharia et al., 2022; Cho et al., 2022; Huang et al., 2023).

Subjectivity A reliable T2I multilinguality assessment must report true knowledge 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, choosing one "correct translation" is in inherently subjective task. This study tackled this subjectivity by casting a wide net of error candidates, and testing their impact. Consequential errors caused *false negatives* where a concept is erroneously marked as "not possessed" (Figure 1).

CCCL's tangible concept constraint and corpus-based approach to finding concepts helps combat subjectivity (Saxon and Wang, 2023). In the tangible sense it's fair to say "orange" is correctly translated in Spanish to *naranja* (the fruit) rather than *anaranjado* (the adjective).

In prompting the T2I model we assume this tangible noun context is induced by using "a picture of an X"-style prompts. While our results show this works, it is a model-specific phenomenon and future work should examine more prompt templates.

Future work grounded in prototype theory (Ando et al., 2002) may enable identification of culturally universal concepts for assessment.

Need to assess Multiple Translations One challenge in multilinguality assessments is *incoming duplicates*, where multiple ways of writing a translation really are equally correct. Our homograph errors have examples, such as cigarette in Japanese. たばこ, タバコ, and 煙草 are all translations of cigarette with identical reading, *tabako*. Why should a metric of model-language capabilities only assess one correct translation rather than all?

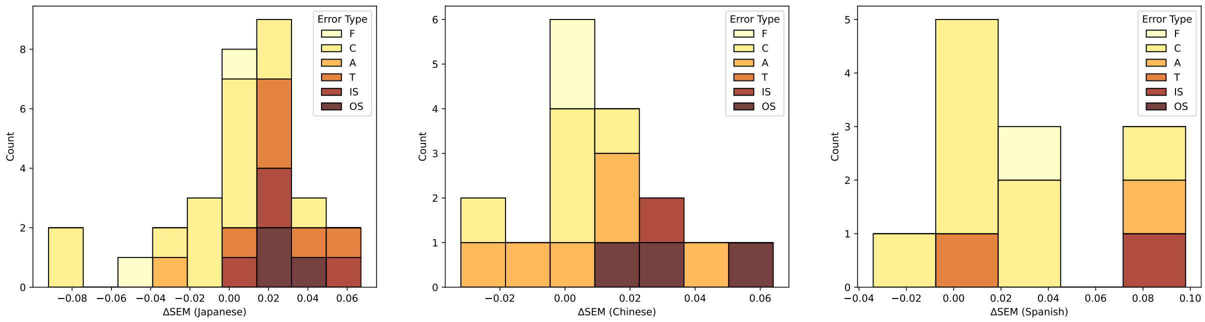


Figure 4: Histograms for the error counts in JA, ZH, and ES vs Δ_{SEM}, colored by error type. From lightest, they are F:*formality*, C:*commonality*, A:*ambiguity*, T:*transliteration*, IS:*incoming sense error*, OS:*outgoing sense error*. The error types are defined in subsection A.3. Severe error types will exhibit more rightward distributional mass.

More significant multiple translation problems arise in languages with gendered human-referent terms. For example, in Spanish *maestro* refers to a male teacher, while *maestra* a female one. Should a test of a model's Spanish knowledge of "teacher" as a concept test that both translations work equally well? CCCL v1 is incapable of assessing these attributes. Future benchmarks should contain this flexibility, so multiple incoming translations (Savoldi et al., 2021) can be assessed for the same concept, while also tracing semantically-encoded secondary attributes such as gender between the source and test language.

Error Severity and Error Type Figure 4 shows the distributions of error types for each language with respect to Δ_{SEM}, our proxy for correction significance or error severity. Across all three languages, the *sense errors* (OS and IS) are the most severe, while the formality and commonality errors are the least severe (defined in subsection A.3).

Our original estimated error rate (sum of all candidates per language) is a worst-case bound, the significant-to-evaluation-validity error rate is lower. Our impact and significance results show that some of our suggestions (mainly formality and commonality errors) may be more nitpick than correction.

Some concepts in CCCL are inherently erroneous due to intangibility. For example, *history*, *film*, and *jump* are all present in v1 of CCCL, picked up for being high-frequency noun concepts across multiple languages in the corpora. There is no sensible prototypical way to generate images "of" those concepts. We removed these for CCCL v1.1; Future benchmarks should avoid including them.

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, X_c doesn't actually change much under this correction as shown in Figure 2 & Table 4. The CLIP similarity score in CCCL is blind to the difference between a bicycle and motorcycle. Mistranslations where visual structural similarity is present are sometimes invisible to the image metrics.

Tangible object translation as an MT domain Single word concepts are not central to the distribution of machine translation training data. By providing the individual English tangible nouns as input we may expect an unreasonable amount of implicit commonsense reasoning from commercial MT systems—the correct sense out of many had to be selected for success. Furthermore, the use of the BabelNet knowledge graph as a consensus mechanism reinforced some sense errors. For example, the *rock* sense error for JA (music genre rather than physical object, Table 4) was also present in Hebrew, probably due to shared edges in the knowledge graph. Given previous interest in assessing the performance of MT translation in diverse domains (Irvine et al., 2013), we think both the word-level translation of concepts under domain constraints without context (as we tried to do in CCCL previously) and treating input prompts for T2I systems (ie, captions) (Hitschler et al., 2016; Singh et al., 2021) as a target domain for MT evaluation would be interesting and useful future directions.

Future benchmarks should leverage context with sentences as input to MT (eg, "watch for falling rocks") rather than the decontextualized concept words alone to improve robustness. LLMs could generate diverse English sentence examples, and could potentially also extract the final concept translations out of the multiple sentence translations.

Limitations

Trivially, human annotators for every language would remove false-negative mistranslations from future benchmarks, but there's a trade-off between easy scalability and certainty of correctness.

Our work incorporates human efforts of both native and proficient but non-native language speakers to propose and resolve translation error candidates 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.

The assumption of *translatability* that underlies CCCL in general is a challenge. As a practical use-based test of functional fairness, using heuristics and only common everyday objects that can be reasonably assumed universal is acceptable, but more linguistic and even philosophical work is needed to really motivate fairness across languages and cultures when underlying assumptions differ.

Acknowledgements

Thanks to our December 2023 ARR reviewers and ACs, particularly Yx9V for thoughtful and detailed reviewing and conversation, and many useful suggestions. Thank you Alfonso Amayuelas for feedback on ES candidates. This work was supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. 1650114, and CAREER Award under Grant No. 2048122.

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.

Maya Ando, Jun Okamoto, and Shun Ishizaki. 2002. Extraction of associative attributes from nouns and quantitative expression of prototype concept. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

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.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2023. Wikiwhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*.

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*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2024. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

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.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models.

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.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

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.

Michael Saxon, Xinyi Wang, Wenda Xu, and William Yang Wang. 2023. PECO: Examining single sentence label leakage in natural language inference datasets through progressive evaluation of cluster outliers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3061–3074, Dubrovnik, Croatia. 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.

Salam Michael Singh, Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021. Multiple captions embellished multilingual multi-modal neural machine translation. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 2–11.

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 Contribution Statement

YL produced the Chinese and Japanese translation error candidates and the overall EC taxonomy. MS produced the Spanish candidates and checked the Japanese candidates. YL evaluated ΔSEM, MS generated the before/after images and evaluated X_c and ΔX_c. YL produced diagrams and MS graphs.

A.1.1 Human Annotation Details

MS and YL produced the initial list of candidate errors and corrections. MS is a native speaker of English and literate second language speaker of Spanish and Japanese. YL is a native speaker of Chinese, professionally proficient speaker of English, and a literate proficient speaker of Japanese, with experience in literary translation and textual localization between English, Chinese, and Japanese.

Each annotator first read through the list of their languages (ES/JA and ZH/JA respectively) for about 10 minutes and marked every translation (*error candidate*) that appeared incorrect with a preliminary correction. They then verified the annotations using bilingual English-{Spanish, Japanese, Chinese} resources and consultation with native speakers where relevant as detailed below.

MS checked Spanish corrections using Spanish-language example usage notes provided in the Spanish wordreference.com dictionary, and consultation with a native speaker. MS's JA error candidates were a subset of YL's. YL also took references from language standard dictionaries used by native speakers—for Chinese *Xiandai Hanyu Cidian* and for Japanese *Shin Meikai Kokugo Jiten*.

A.2 Additional Resource Information

Intended Use, License and Terms We release our corrections as a v1.1 revision to the CoCo-CroLa benchmark (Saxon and Wang, 2023) intended to evaluate the performance of text-to-image models. It inherits v1's license and terms.

Offensive Content Some of the erroneous translations we found can lead to offensive images, e.g. the original JA translation for *milk* in also means "breast."

A.3 Error candidate typology

Commonality (C). When a selected translated term doesn't appear to reflect the most common, colloquial, contemporary, or "natural" way that native speakers of the language would use in reference to the concept in a photograph or conversation. For example, in Chinese "瓶子" is a more conversational and contemporary way of writing *bottle* than "瓶," which reads literary and archaic.

Outgoing Sense Error. (OS) 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.

Incoming Sense Error. (IS) The translated term, while aligned to the correct source concept sense, picks a phrasing for which other senses in the target language exist that the annotators expect will confound model behavior, where another (often more common) disambiguated translation also exists. For example, the original Spanish translation for *tent* is given as *tienda* alone, which can also mean 'store, shop', in addition to 'a tent,' whereas the corrected translation *tienda de acampar* refers to a camping tent alone.

Ambiguity (A). 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. (F) The translated term uses an expression in an improper formality. For example, the original Chinese translation for *Father* is only heard in casual conversations.

Transliteration (T). When one of the above errors occurs with . For example, the transliteration of *Rock* in Japanese is commonly related to 'Rock Music', rather than stones found in nature.

A.4 Computational Experiments Details

Dataset Statistics CCCL contains 193 multilingual concepts written in 7 languages. We have also modified 50 of these in ES, ZH, or JA with verified translations by human annotators.

Models Employed See Table 2.

Model	# Param	Repository (huggingface.co/...)
StableDiffusion 1.4	860M	`CompVis/stable-diffusion-v1-4`
StableDiffusion 2	NA	`stabilityai/stable-diffusion-2`
StableDiffusion 2.1	NA	`stabilityai/stable-diffusion-2`
AltDiffusion m9	1.7B	`BAAI/AltDiffusion-m9`

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

Experimental Setup We generated 9 images for each (language, model, concept) triple and evaluated X_C using identical methods and codeas described in CCCL (Saxon and Wang, 2023).

A.5 Full Analysis Numbers

Model	Language	PCC	p	m	b
SD1-4	Japanese	0.120	0.577	0.437	0.049
SD1-4	Chinese	0.018	0.944	0.051	-0.011
SD1-4	Spanish	0.384	0.307	1.877	-0.064
SD2	Japanese	0.088	0.684	0.155	0.020
SD2	Chinese	0.155	0.554	0.608	0.000
SD2	Spanish	0.646	0.060	3.891	-0.067
AD	Japanese	0.734	0.000	1.519	0.014
AD	Chinese	0.725	0.001	4.472	-0.010
AD	Spanish	0.895	0.001	3.588	0.010
SD2-1	Japanese	0.162	0.448	0.272	0.013
SD2-1	Chinese	0.078	0.765	0.340	0.001
SD2-1	Spanish	0.574	0.106	3.722	-0.075

Table 3: Stats for Pearson correlation and linear best fit between ΔSEM and ΔX_c for each model and language. p represents the p-value for the PCC, m and b the slope and intercept for the best-fit line.

A.6 Further Related Work

ConceptBed (Patel et al., 2024) evaluates monolingual concept-level knowledge in T2I, and its concept inventory could extend and improve CCCL's. T2I-CompBench assesses compositionality in T2I (Huang et al., 2024), leveraging VQA and image segmentation. Assessment model weaknesses, such as Agrawal et al. (2018)'s VQA spurious correlations (Antol et al., 2015) remain a challenge.

Other benchmarks in vision-and-language also require correction and improvement. Luo et al. (2022) found and filtered unsolvable cases in *Who's Waldo* (Cui et al., 2021). Ye and Kovashka (2021) exploit repeated texts in QA pairs on *VCR* (Zellers et al., 2019). While manual techniques can find and clean these errors, automated approaches would be preferable, such as the PECO method (Saxon et al., 2023) for finding model-used shortcuts in NLI. Semi-human-in-the-loop approaches (Ho et al., 2023) may improve the sourcing and cleaning of future CCCL versions.

Figure 5: 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.

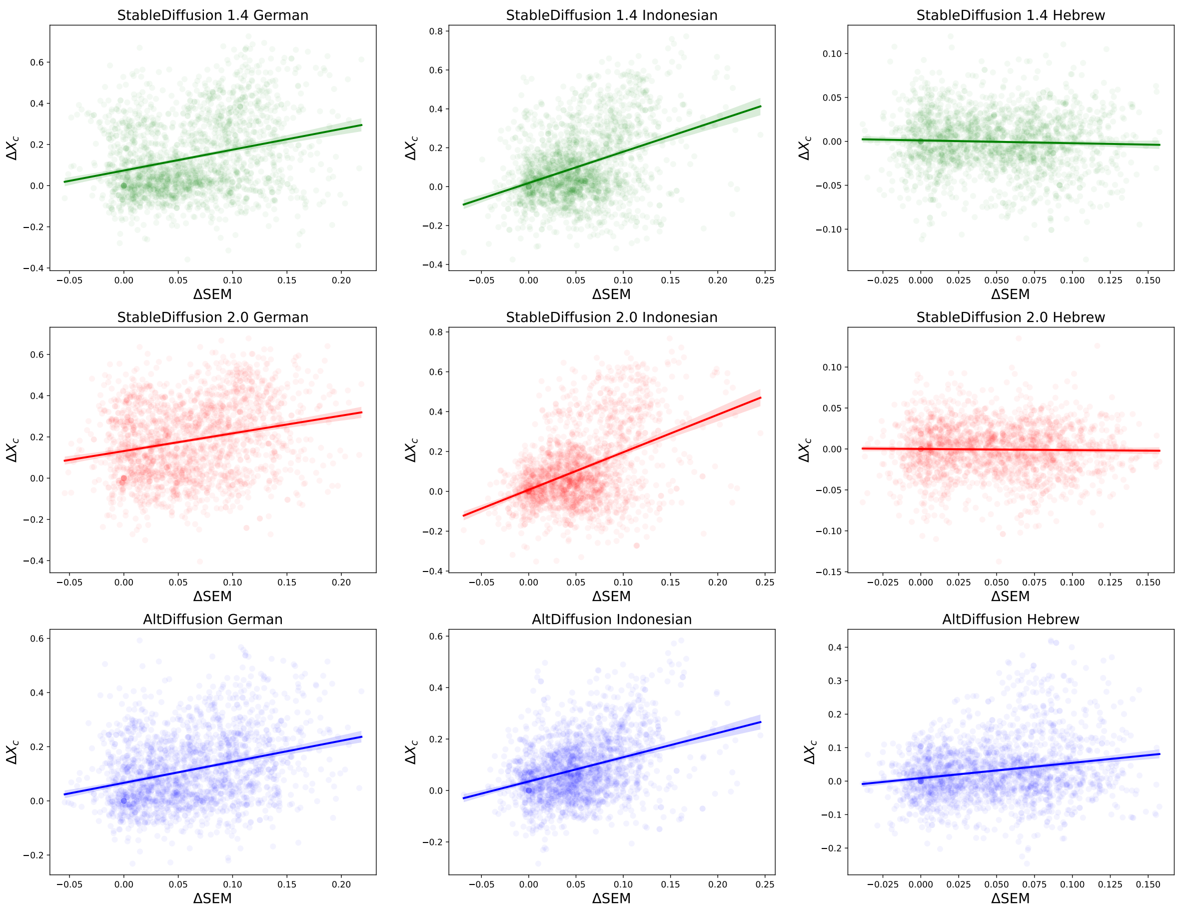


Figure 6: Scatterplots for the pseudocorrection experiments. Transparent circles are used to make distribution mass more visible.

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

Table 4: 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. Error types are defined in subsection A.3