Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset

Anonymous ACL submission

Abstract

Research in massively multilingual image captioning has been severely hampered by a lack of high-quality evaluation datasets. In this paper we present and make available the Crossmodal-3600 dataset, a geographically-diverse set of 3600 images each of them annotated with human-generated reference captions in 36 languages. We select a representative set of images from across the world for this dataset, and annotate it with captions that achieve consistency in terms of style across all languages, while avoiding annotation artifacts due to direct translation. We apply this benchmark to model selection for massively multilingual image captioning models, and show superior correlation results with human evaluations when using the Crossmodal-3600 dataset as golden references for automatic metrics.

1 Introduction

Image captioning consists in automatically generating a fluent natural language description of a given image. This task is important for enabling accessibility for visually impaired users, and is a core task in multimodal research encompassing both vision and language modeling. However, datasets for this task are primarily available in English (Young et al., 2014; Chen et al., 2015a; Krishna et al., 2017; Sharma et al., 2018; Pont-Tuset et al., 2020). Beyond English, there are a few datasets such as Multi30K with captions in German (Elliott et al., 2016), French (Elliott et al., 2017) and Czech (Barrault et al., 2018), but they are limited to a few languages that cover a small fraction of the world's population and feature images that severely underrepresent the richness of cultures from across the globe. These aspects have hindered research on image captioning for a wide variety of languages, and directly hamper deploying accessibility solutions for a wider audience of visually impaired people from around the world.

Creating large training and evaluation datasets in several languages is a resource intensive endeavor, but recent works (Thapliyal and Soricut, 2020) have shown that it is feasible to build multilingual image captioning models trained on machine-translated data (with English captions as the starting point). But they have also shown that the effectiveness of some of the most reliable automatic metrics for image captioning, such as CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016), is severely diminished when applied to translated evaluation sets, resulting in even poorer agreement with human evaluations compared to the English case. As such, the current situation is that trustworthy model evaluation can only be based on extensive side-by-side human evaluations, but such evaluations cannot usually be replicated across efforts and do not offer a fast and robust mechanism for model hill-climbing and comparison of multiple lines of research work.

The Crossmodal-3600 image captioning evaluation dataset provides a robust benchmark for multilingual image captioning and can be reliably used to compare research contributions in this emerging field. Our contributions are as follows: (i) for caption annotations, we have devised a protocol that allows human annotators for a specific target language to produce image captions in a style that is consistent across languages, for all 36 languages we considered, and with multiple replication; moreover, this protocol facilitates image-caption creation that is free of direct translation artefacts, an issue that has plagued Machine Translation research for many years and it is now well understood (Freitag et al., 2020); (ii) for image selection, we have devised an algorithmic approach to sample a set of 3600 geographically-diverse images from the Open Images Dataset (Kuznetsova et al., 2020), aimed at creating a representative set of images from across the world; (iii) for the resulting benchmark, we empirically measure its ability to rank image cap-

Source: 20150726_120207 by Nguyen Hung Vu.

English	• A macro shot of pink flower in the garden • A close-up view of the pink flower with green leaves in garden
French	• Des fleurs de bégonia dans un jardin *(Begonia flowers in a garden)* • Fleurs rose en gros plan avec en arrière plan plusieurs autres flou *(Pink flowers in close-up with in the background several others blur)*
Hindi	• गुलाबी रंग के फूल का करीबी दृश्य है और पृष्ठभूमि में पौधे का धुंधला नज़ारा है *(There is a close view of the pink flower and the background is a blurred view of the plant)* • धुंधले प्रष्ठभूमि में गुलाबी रंग के फूलो का करीबी दृश्य *(Close-up view of pink flowers in the misty landscape)*

Figure 1: Sample captions in three different languages (out of 36 – see full list in Appendix A), showcasing the creation of annotations that are consistent in style across languages, while being free of direct-translation artefacts (e.g., the French version with "bégonia" would not be possible when directly translating from the English versions).

tioning model variations such that it provides high agreement with human judgements, therefore validating its usefulness as a benchmark and alleviating the need for human judgement in future research.

Fig. 1 shows a few sample captions for an image in the dataset that exemplify point (i) above, and Fig. 2 shows the variety of cultural aspects captured by the image sampling approach from point (ii). We provide detailed explanations and results for each of the points above in the rest of the paper. We are releasing the Crossmodal-3600 dataset under a CC-BY4.0 license.

2 The Crossmodal-3600 Dataset

2.1 Language Selection

As a first step, we take a quantitative stance for the language-selection problem and choose 30 languages roughly based on their percent of web content; we call this set of languages L30[1]. As a second step, we consider an additional five languages (L5[2]) to cover low-resource languages with many native speakers, or major native languages from continents that would not be covered otherwise. The protocol for caption annotation (Sec. 2.3) has been applied for the resulting union of language plus English, for a total of 36 languages.

2.2 Image Selection

For each of the 36 languages we target, we select 100 images that, as far as it is possible for us to identify, are taken in an area where the given language is spoken. The images are selected among those in the validation and test splits of the Open Images Dataset (Kuznetsova et al., 2020) that have GPS coordinates stored in their EXIF metadata.

Since there are many regions where more than one language is spoken, and given that some areas are not well covered by Open Images, we design an algorithm that maximizes the percentage of selected images taken in an area in which the assigned language is spoken. This is a greedy algorithm that starts the selection of images by the languages for which we have the smallest pool (e.g. Persian) and processes them in increasing order of their candidate image pool size. Whenever there are not enough images in the area where a language is spoken, we have several back-off levels: (i) selecting from a country where the language is spoken; (ii) a continent where the language is spoken, and, as last resort, (iii) from anywhere in the world.

This strategy results in 100 images from an appropriate region for most of the 36 languages except for Persian, where 14 continent-level images are used, and Hindi, where all 100 images are at the global level because the in-region images are assigned to Bengali and Telugu. We keep the region each image is selected from as part of our data annotation so that future evaluations can choose to only evaluate on images relevant to particular regions of interest, or on the entire dataset.

2.3 Caption Annotation

For a massively multilingual benchmark such as the one we created, consistency in the style of the description language is critical, since language can serve multiple communication goals. For a more in-depth discussion on these issues as they relate to image captions, we refer the reader to (Alikhani et al., 2020). We borrow from their terminology re-

[1] French (fr), Italian (it), German (de), Spanish (es), Hindi (hi), Arabic (ar), Chinese-Simplified (zh), Dutch (nl), Japanese (ja), Korean (ko), Polish (pl), Portuguese (pt), Russian (ru), Thai (th), Turkish (tr), Croatian (hr), Czech (cs), Danish (da), Finnish (fi), Greek (el), Hebrew (iw), Hungarian (hu), Indonesian (id), Norwegian (no), Romanian (ro), Vietnamese (vi), Farsi (fa), Swedish (sv), Ukrainian (uk), Filipino (fil).

[2] Swahili (sw), Maori (mi), Cusco Quechua (qu), Telugu (te), Bengali (bn).

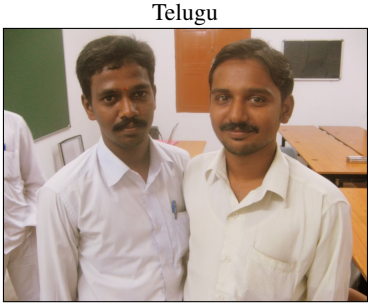
English

Source: 2011 Lake Mirror Classic Auto Festival by Catch Central Florida

Swahili

Source: Kenyan TED Fellows Dinner. High fives and happiness all around. by afromusing

Telugu



Source: Farewell to Dinesh by Ramana

Quechua

Source: Pisac - Mercado by pululante

Filipino

Source: Cebu-Manila-Baguio by Muiz Misri

Chinese

Source: 2004-11-29 12-02 Singapur 036 by Allie_Caulfield

Figure 2: A sample of images in the Crossmodal-3600 dataset, together with the language for which they have been selected. The images span regions over six different languages and over four different continents.

lated to the coherence relations between image and captions: VISIBLE, META, SUBJECTIVE, STORY relations. The goal for our caption annotation is to generate VISIBLE image captions, i.e., presenting information that is intended to recognizably characterize what is depicted in the image.

One possible approach to generating captions that achieve this goal is to generate them as such in English, and have them translated (automatically, semi-automatically, or manually) into all the other languages. However, this approach results in an English-language bias, as well as other problems that have been already identified in the literature. For instance, translations are often less fluent compared to natural target sentences, due to word order and lexical choices influenced by the source language. The impact of this phenomenon on metrics and modeling has recently received increased attention in the evaluation literature (Toral et al., 2018; Zhang and Toral, 2019; Freitag et al., 2020), and references created in this style are thought to cause overlap-based metrics to favor model outputs that use such unnatural language.

We have designed our caption annotation process to achieve two main goals: (i) produce caption annotations in a VISIBLE relation with respect to the image content, and, strongly, create consistency in the description style across languages; (ii) be free of translation artefacts. To achieve this, we use bi-lingual annotators with a requirement to be reading-proficient in English and fluent/native in the target language. As a preliminary step, we train an image-captioning model on English-annotated data, which results in captions in the VISIBLE style of COCO-CAP (Chen et al., 2015b).

The annotation process proceeds as follows. Each annotation session is done over batches of $N = 15$ images, using the images selected as described in Sec. 2.2. The first screen shows the N images with their captions in English as generated by the captioning model, and asks the annotators if the captions are EXCELLENT, GOOD, MEDIUM, BAD, or there is NOT-ENOUGH-INFO. We provide the annotators with clear guidelines about what constitutes an EXCELLENT caption, and how to evaluate degradations from that quality. This step forces the annotators to carefully assess caption quality and it primes them into internalizing the style of the captions without the need for complicated and lengthy annotation instructions.

The second screen shows the same N images again but without the English captions, and the annotators are asked to produce descriptive captions in the target language for each image. In the absence of the English captions, the annotators rely on their internalized caption style, and generate their annotations mostly based on the image content – with no support from the text modality, other than potentially from memory. Note, however, that we have designed the system to support N anno-

tations simultaneously, and we have empirically selected the value of N as to be large enough to "overwrite" the memory of the annotators with respect to the exact textual formulation of the English captions. As a result, we observe that the produced annotations are free of translation artefacts, see the example in Fig. 1 for French mentioning "bégonia", and for Hindi mentioning "misty landscape".

We also provided the annotators with a systematic heuristic to use when generating captions, which provided useful guidance in achieving consistent annotations across all the language targeted. We provide the annotations guidelines in Appendices B and C. For each language, we generated captions over all 3600 images with replication 2 (two different annotators working independently)[3], except Bengali (bn) with replication 1.

2.4 Caption Statistics

Table 1 provides detailed caption statistics, including number of distinct captions per image and average words and characters per caption. There are a total of 256,990 distinct captions across 36 languages, each image having in the vast majority of cases at least 2 distinct captions per language.

For languages with natural space tokenization, the numbers of words per caption can be as low as 5 or 6 for some agglutinative languages like Quechua (qu) and Czech (cs), and as high as 18 for an analytic language like Vietnamese (vi). The number of characters per captions also varies drastically – from mid-20s for Korean (ko) to mid-90s for Indonesian (id) – depending on the alphabet and the script of the language.

2.5 Caption Quality

To ensure quality, the annotation process is initially started with pilot runs on 150 images until very few low-quality captions are being produced[4]. Then we run the main annotation and finally a verification round where we select one caption for 600 randomly selected images and have the annotator pool (per language) rate them on the same quality scale used in the experiment: EXCELLENT, GOOD, MEDIUM, BAD, and NOT-ENOUGH-INFO. The quality scores are presented in Table 2.

[3]Due to various issues related to process idiosyncrasies, the exact replication varies slightly under or over 2 (see Tab. 1). Maori (mi) annotations are currently missing from the table, and we plan to add them as soon as they become available.

[4]Between one and five pilots were needed per language

Lan. Id.	Num. Cap.	Replication 1	Replication 2	Replication 3+	Num. Words	Num. Chars
ar	7362	4	3431	165	7.8	42.3
bn	3600	3600	0	0	11.3	62.1
cs	7061	157	3432	11	6.5	39.1
da	7264	0	3542	58	8.7	48.3
de	8643	0	2240	1360	11.2	76.5
el	7202	2	3594	4	7.8	51.4
en	8527	144	2030	1426	12.0	61.0
es	8614	0	2201	1399	9.8	56.3
fa	7245	0	3555	45	12.8	59.4
fi	7126	90	3500	10	7.5	65.1
fil	7123	77	3523	0	12.2	67.4
fr	8562	0	2253	1347	12.3	69.5
hi	8502	0	2298	1302	13.4	59.9
hr	7276	2	3551	47	9.0	57.8
hu	7215	1	3585	14	8.6	60.5
id	7126	74	3526	0	14.3	93.5
it	8471	0	2329	1271	12.1	71.8
iw	7200	0	3600	0	11.9	63.6
ja	7185	15	3585	0	1.0	26.0
ko	7649	16	3314	270	7.0	24.7
nl	7312	3	3507	90	9.2	53.0
no	7195	5	3595	0	9.6	54.3
pl	7140	60	3540	0	8.4	57.6
pt	7243	0	3562	38	10.8	61.7
qu	7200	0	3600	0	5.0	38.7
ro	7123	77	3523	0	15.6	88.4
ru	7200	0	3600	0	9.9	66.3
sv	7273	1	3536	63	8.1	46.7
sw	7046	154	3446	0	10.7	63.1
te	7200	0	3600	0	7.1	47.5
th	7200	0	3600	0	1.2	47.9
tr	7231	17	3536	47	9.4	63.4
uk	7214	1	3584	15	10.0	65.7
vi	7350	0	3450	150	18.0	79.3
zh	7110	90	3510	0	1.0	23.0

Table 1: **Caption statistics**: A total of 256,990 distinct captions across 36 languages.

2.6 Annotator Details

We use an in-house annotation platform with professional (paid) annotators and quality assurance. Annotators are chosen to be native in the target language whenever possible, and fluent otherwise (for low-resource languages, they are usually linguists that have advanced-level knowledge of that language). All annotators are required to be proficient in English since the instructions and guidelines are in English.

3 Using Crossmodal-3600 for Model Comparison

In this section we compare different flavors of multilingual image captioning models on Crossmodal-3600. As first baseline we use TGT (i.e., an

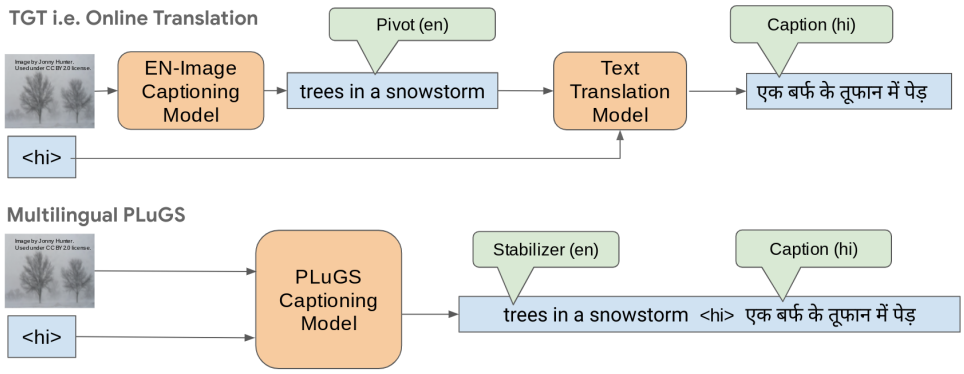


Figure 3: A comparison between the TGT (on-line translation) and PLuGS methods for generating non-English captions, starting from English training data. The PLuGS method is superior because the Caption (target-language) ends up being a translation of the Stabilizer (en) done in the presence of the image input.

Lang Id	%GOOD+	%MED+	%BAD
ar	97.5	99.3	0.7
bn	100.0	100.0	0.0
cs	96.8	99.0	1.0
da	94.0	99.2	0.8
de	98.2	99.3	0.7
en	95.5	98.3	1.7
es	97.0	98.3	1.7
fa	94.0	99.3	0.7
fil	79.7	95.3	4.5
fr	92.7	99.2	0.8
hi	92.7	98.7	1.3
hr	80.7	98.2	1.8
hu	91.3	94.8	5.0
it	88.8	97.7	2.3
iw	82.7	96.7	3.0
ja	84.3	96.3	3.5
pl	92.2	97.3	2.7
pt	87.8	99.5	0.3
ro	90.2	98.3	1.7
ru	93.8	99.5	0.3
te	98.7	99.8	0.2
tr	97.8	98.0	1.2
uk	91.2	99.2	0.8
vi	94.3	97.8	2.0
zh	90.2	97.8	2.2

Table 2: **Caption quality statistics** for 25 of the 36 languages, we show the median of three ratings (ar and fil have only one rating).

English-model captioning followed by on-line translation, in this case using the Google Translate API), see Fig. 3 (top). The other models are all multilingually trained over 31 languages (English + L30), using translated captions as training material (again, using Google's API). As our main result, we show superior correlation between model rankings using human-evaluation scores for these models and scores obtained using CIDEr (Vedantam et al., 2015) with the Crossmodal-3600 dataset for gold-caption references.

3.1 PluGS Models

We start from the approach proposed in (Thapliyal and Soricut, 2020) and train a multilingual image captioning model for English + L30 languages, using the PLuGS (Pivot Language Generation Stabilized models) architecture. The backbone model is a Transformer encoder-decoder network (Vaswani et al., 2017). The input sequence for the encoder consists of the following:

1. **Global Image Representation**: A global image representation embedding (of dimension 64) from Graph-RISE (Juan et al., 2019), based on a ResNet-101 (He et al., 2016), and projected to the internal Transformer dimension (512 in our experiments) by a 2-layer DNN with linear activation.
2. **Image Objects Representations**: A ResNet-101 object-detection classifier trained on JFT (Hinton et al., 2015) produces a list of detected object-label identifiers, sorted in decreasing order by the classifier's confidence score; the first sixteen are mapped to an object embedding using an object-label embedding layer pre-trained using the word2vec approach (Mikolov et al., 2013) to predict label co-occurrences in web documents; similar to the global-image embedding, these embeddings are projected to match the Transformer internal dimension.
3. **Language Identifier (LangId)**: The target language using an identifier string, encoded with a LangId vocabulary (0-30) and projected using a 2-layer DNN to match Transformer dimensions and produce a LangId embedding.

Stabilizer	Occurrences
example of a trendy bedroom design with gray walls	15
example of a trendy bedroom design	6
example of a minimalist bedroom design	4
example of a trendy dark wood floor bedroom design with gray walls	3
example of a minimalist bedroom design with gray walls	2
example of an eclectic bedroom design with gray walls	1

Table 3: Examples of inconsistent stabilizers (and hence caption meanings)

Both input and output text are byte-pair encoded (Sennrich et al., 2016) with a shared source-target vocabulary of 12,000 tokens, mapped to a sequence of text embeddings with dimensions matching Transformer dimensions. We have reserved token-ids for each language (e.g. $\langle de \rangle$ for German) used as separators in the PLuGS target output.

The defining feature of PLuGS models is that first the caption in the pivot language (en) is generated (called stabilizer), then the target language separator (e.g. $\langle de \rangle$), followed by the caption in the target language (see Fig. 3). The intuition behind this design is that it is advantageous for learning to allow the Transformer decoder to self-attend on the gold-data English caption (as a result of teacher-forcing) when learning to predict the non-English caption tokens, and our empirical results (see Table 5, top 10 rows) support this intuition.

To get the final caption, the model output is split on the separator to obtain two strings, the stabilizer caption and the target-language caption (Fig. 3, bottom). We note that the caption generated by a PLuGS model tends to be a translation (Thapliyal and Soricut, 2020) of the stabilizer done *in the context of the image input* and therefore superior to performing a text-based translation.

3.2 Consistent PLuGS Models

While we have replicated and observed the advantages of the PLuGS model over direct translation (TGT), there is an undesirable side-effect to this method: for the same image but different target languages (say, French and Hindi), the generated stabilizers can be different, and – since the target-language caption tends to be a translation of the stabilizer – the two caption outputs can be semantically different. We exemplify this phenomenon in Table 3 with the output produced by our PLuGS En+L30 model, and note that the semantics of the generated captions can differ significantly. We also measure quantitatively the severity of this phenomenon in Fig. 4, showing the distribution

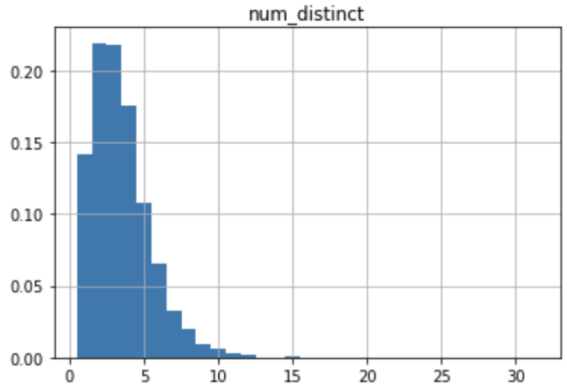


Figure 4: **Normalized frequency** of number of distinct stabilizers for the En+L30 PLuGS model. The mode is around 3-4 distinct stabilizers per image, which may result in significant differences in the semantics of the generated captions across various languages.

in the number of distinct stabilizers (normalized over the development set) produced by the PLuGS En+L30 model. Since the mode of this distribution is around 3-4 distinct stabilizers, this means that it is quite likely that the target-captions produced – for the same image – for different target languages are not semantically equivalent. To fix this issue, we propose several small modifications to the PLuGS model, yielding what we call a Consistent-PLuGS model. This is achieved with the following changes.

First, at both training and inference time, instead of feeding the $\langle$LangId$\rangle$ token to the encoder, we feed it instead to the decoder – after it has generated the stabilizer. As a result, the generation of the stabilizer is independent of the target language. In order to be able to determine when the stabilizer generation has ended, we introduce an additional end-of-stabilizer $\langle$EOSTAB$\rangle$ token.

Second, at inference time, we note that the model cannot predict the $\langle$LangId$\rangle$ token after $\langle$EOSTAB$\rangle$ since, by design, it is never been given that information up to that point. Therefore, we use a two-step decoding process: first we decode until $\langle$EOSTAB$\rangle$,

then we insert the target ⟨LangId⟩ token into the output and continue with the subsequent decoding.

Further, under beam search, this procedure achieves language-independent decoding only for a beam of size 1. When decoding with a beam size > 1, a beam search over the entire output sequence (except for the ⟨LangId⟩ token) still results in "inconsistent decoding" (due to the coupling between the stabilizer and the target-language caption, we still observe different stabilizers for different languages, albeit with a mode shifted to the left compared to Fig. 4). Thus, in this case we do a two-phase decoding: in the first phase we produce the top stabilizer; in the second phase, we hold the stabilizer fixed and perform beam search over the rest of the sequence after the ⟨LangId⟩ token. We call this procedure "consistent decoding", and refer to the resulting outputs as being produced by a Consistent-PLuGS model.

3.3 Training Details

Our models are trained using the Conceptual Captions 3M dataset (Sharma et al., 2018), translated to the L30 languages using Google's machine translation API. We use the standard train and validation splits provided with the dataset[5]. The models are trained on a 4x4 TPU architecture. The English models are trained using a Stochastic Gradient Descent optimizer with a linear warm-up of 16k, and base learning rates and corresponding halving decay steps scanned over {(0.18, 80k), (0.18, 130k), (0.18, 200k), (0.24, 70k), (0.24, 100k), (0.32, 50k), (0.32, 70k)}. A vocabulary size of 4k which worked as well as larger sizes. The PLUGS models are trained using the Adam optimizer (Kingma and Ba, 2015) with a linear warm-up of 16k and base learning rates and corresponding halving decay steps scanned over {(1e−4, 100k), (1e−4, 150k), (2e−4, 50k), (2e−4, 100k)}. Furthermore, a dropout of 0.3 and L2 regularization weight of 1e−5 is used for all the trainable parameters for all models. For PLuGS models, we experimented with various vocabulary sizes, and found that 12k worked as well as higher sizes for our En+L30 setup. We also observed that oversampling the English captions when creating the vocabulary models sometimes gave us significant gains in performance, so we ran each PLuGS training with English oversampling factors of 1, 30 and 90.

Table 4 describes the best models we found in each class, and gives details for each model configuration used in our quantitative experiments. The first five models in the table have around 57 million parameters while L31-CMCD8 has around 65 million parameters and L31-CMCD10 has around 73 million parameters. Together, all the experiments we conducted took around 123k GPU hours.

3.4 Human Evaluation

To be able to create a gold reference for the quality of the models from Table 4, we conduct side-by-side human evaluations using the outputs of these models. In order to simulate a more realistic scenario, we use a set of 1000 randomly selected images from the Open Images Dataset, distinct from the images used in the Crossmodal-3600 dataset. Image captions generated by a given pairing of models (model1 vs model2, where model1 is considered as the base condition and model2 the test condition) are compared and rated side-by-side, using a similar pool of raters as described in Sec. 2.6. Each side-by-side pair (shown in a random order for each example) is rated using a scale of MUCH-BETTER, BETTER, SIMILAR, WORSE, MUCH-WORSE, with a replication factor of 3.

We use a side-by-side metric defined using the following values: WINS = % of images where a majority of raters (i.e. 2 out of 3) mark model2 captions as better; LOSSES = % of images where majority of raters mark model2 captions as worse; the overall metric defined as SXSGAIN = WINS - LOSSES.

3.5 Results

In Table 5, we present the results of model comparisons between various model pairs, including comparisons between PLuGS model (both regular and consistent) and TGT, and between different PLuGS model variants. The goal of this comparison is to determine which model variants are performing the best. Note that this setup is both realistic (mimicking real-life situations in which close-variant models need to be evaluated against each other) and more difficult than scenarios in which models with significantly different architectures are compared. The gold-reference for the relative strength of each pairing is given by the SxSGain column, with positive numbers indicating the superiority of the Model2 variant, and negative numbers indicating a superiority for the Model1 variant. The 3600_diff column is capturing similar information,

[5]Train: 3,318,333 image-caption pairs. Validation: 15,840 image-caption pairs.

Model	Details	Parameters
TGT	Machine-translated captions generated by the English model.	*enc=6, dec=6, lr= 0.18 (SGD), dc=200k, bs=4k, v=4k*
L31	Plain PluGS model with beam search decoding	*oos=1, enc=6, dec=6, lr=1e−4, dc=150k, bs=4k, v=12k*
L31-CMCD	Consist. PluGS model and consistent beam search decoding	*os=1, enc=6, dec=6, lr=1e-4, dc=150k, bs=4k, v=12k*
L31-CMID	Consist. PluGS model and inconsistent beam search decoding	*os=1, enc=6, dec=6, lr=1e-4, dc=150k, bs=4k, v=12k*
L31-CMGD	Consist. PluGS model and greedy decoding	*os=1, enc=6, dec=6, lr=1e-4, dc=150k, bs=4k, v=12k*
L31-CMCD8	Consist. PluGS model and consistent beam search decoding	*os=30, enc=6, dec=8, lr=1e-4, dc=150k, bs=8k, v=12k*
L31-CMCD10	PluGS with decoder-side translation and consistent beam search	*os=30, enc=6, dec=10, lr=1e-4, dc=150k, bs=8k, v=12k*

Table 4: **Model details** for all model variants used in our experiments: *os* denotes the English oversampling factor when creating the vocabulary; *enc*/*dec* denote the number of encoder/decoder layers; *bs* denotes the training batch size; *v* denotes the size of the vocabulary; *dc* denotes the number of steps over which the learning rate is halved.

Model 2	Model 1	Lang. Id	SxS Gain	ΔCIDEr (3600)	ΔCIDEr (Val)
L31-CMCD	TGT	en	2.70	2.61	−2.78
L31-CMCD	TGT	fr	3.00	0.84	−5.64
L31-CMCD	TGT	hi	2.20	1.37	−4.38
L31-CMCD	TGT	es	2.70	1.96	−5.55
L31-CMCD8	TGT	en	−0.30	1.28	1.56
L31-CMCD8	TGT	hi	1.60	1.16	−1.71
L31-CMCD8	TGT	es	3.10	1.22	−2.38
L31-CMCD10	TGT	en	4.30	−0.44	3.51
L31-CMCD10	TGT	hi	0.80	1.16	0.00
L31-CMCD10	TGT	es	2.70	0.20	−0.50
L31-CMCD	L31	en	−0.30	−0.33	−0.64
L31-CMCD	L31	hi	−1.70	−0.39	−1.03
L31-CMCD	L31	es	−2.20	−0.51	−1.10
L31-CMCD	L31-CMG	en	6.00	0.96	3.53
L31-CMCD	L31-CMG	hi	1.90	0.47	3.07
L31-CMCD	L31-CMG	es	4.00	1.12	2.99
L31-CMCD	L31-CMG	hi	−0.20	−0.10	−1.10
L31-CMCD	L31-CMG	es	−0.60	−0.45	−1.31

Table 5: **Model comparison** (Model 2 vs Model 1). *Lang* denotes the target language; ΔCIDEr(3600) is CIDEr(Model 2)-CIDEr(Model 1) on the Crossmodal-3600 dataset, while ΔCIDEr(Val) on the validation split with machine-translated references.

except its numbers are based on CIDEr scores using the Crossmodal-3600 dataset as references, while the val_diff column is using numbers based on using machine-translated references obtained from the validation split of CC3M.

The numbers presented in Table 5 are used to compute the correlation between human judgements regarding the relative quality of the captioning models and the ability of the CIDEr[6] metric – or, rather, of the underlying references used by the metric – to perform a similar judgement.

The correlation results are presented in Table 6, using various correlation formulations. The comparison between the performance obtained using the Crossmodal-3600 dataset and the translated validation dataset clearly indicate large improvements in correlation from using the Crossmodal-3600 references, with all four correlation metrics supporting this conclusion. Based on these results, we recommend the use of the Crossmodal-3600 references as a superior way to quantify and judge relative model strengths.

[6]We used the reference impl. with default parameters: github.com/vrama91/cider. We multiply the score by 100.

Correlation Coefficient	ΔCIDEr Crossmodal	ΔCIDEr Validation
Pearson	0.47 (medium)	0.19 (weak)
Spearman	0.38 (medium)	0.12 (weak)
Kendall	0.31 (strong)	0.08 (weak)
Matthews	0.75 (strong)	0.18 (weak)

Table 6: **Correlation** between human evaluations and CIDEr difference on Crossmodal-3600 and the val. set.

4 Conclusions

We introduce the Crossmodal-3600 dataset as a benchmark for evaluating the performance of multilingual image captioning models. The images in the dataset are geographically diverse, covering all inhabited continents and a large fraction of the world population. We believe this benchmark has the potential to positively impact both the research and the applications of this technology, and enable (among other things) better accessibility for visually-impaired users across the world, including speakers of low-resource languages.

The main appeal of this benchmark is that it alleviates the need for extensive human evaluation, which is difficult to achieve across multiple languages and hinders direct comparison between different research ideas and results. We show significant improvements in correlation with human judgements when using the Crossmodal-3600 dataset as references for automatic metrics, and therefore hope that the adoption of this dataset as a standard benchmark will facilitate faster progress and better comparisons among competing ideas.

5 Ethical Considerations

5.1 Risks

The approach to data collection of CC3M (Sharma et al., 2018) upholds rigorous privacy and ethics standards, such as the removal of offensive content, personal identification data, and hypernymization. This significantly mitigates the risks that the captioning models we train would produce such information. Similarly, the Crossmodal-3600 dataset is free of such risks, as the annotations has been produced in-house and have been quality controlled, while the images used have been vetted to be appropriate for the intended use.

5.2 Limitations

Due to the high volume of work required and the cost associated with it, we have only targeted 36 languages for our annotation effort; while this number is significantly higher than what is available with previous annotations, it still falls short of including many other languages spoken and written around the world.

For the same reason mentioned above, we have sampled only 100 images for each of the targeted languages, which limits the amount of natural and cultural phenomena that these images capture. While the resulting 3600 images have significantly more variety compared to previous datasets, it may still fall short of including important aspects of natural and cultural life from around the globe.

References

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *ECCV*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015a. Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015b. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *ArXiv*, abs/2004.06063.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-rise: Graph-regularized image semantic embedding. *CoRR*, abs/1902.10814.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 123(1):32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the ACL*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Ashish V. Thapliyal and Radu Soricut. 2020. Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 160–170, Online. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. 2:67–78.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Additional Caption Examples

Figure 5 displays the captions in the 36 languages covered in Crossmodal-3600 for the same image as in Figure 1.

B Instructions for Rating Captions

The following instructions are provided to annotators for rating captions:

This task involves rating captions. To guide your ratings, imagine that you are describing the image to a visually impaired friend, then consider: how well does the caption describe the image to this friend?

Use the following scale for judging the quality of the captions (for borderline cases, use the lower rating):

- BAD: The caption has one or more of the following issues: a). Caption misses the main topic of the image. b). Caption has major grammatical errors (such as being incomplete, words in wrong order, etc). Please ignore capitalization of words and punctuation. c). Caption violates the 'No Hallucination' rule by mentioning objects, activities, or relationships that are definitely not in the image. Note: Apply the 'No-Hallucination' rule only when you are certain that an object/activity/relationship is definitely not implied by the image (see the examples below).
- MEDIOCRE: The caption may capture some objects and activities but misses crucial information (related to activity, important objects/persons in the scene, important modifiers, etc.)
- GOOD: The caption explains most of the main objects, activities, and their relationships in the image.
- EXCELLENT: The caption covers well the whole image, including all the main objects, activities, and their relationships.
- NOT ENOUGH INFORMATION: Not enough information to evaluate the caption quality. Please try to use one of the four categories above as much as possible. Assume that any missing information is favorable to the caption rather than against it.

C Instructions for Generating Captions

The following instructions are provided to annotators for generating captions:

To guide your caption generation, imagine that you are describing the image to a visually impaired friend. The caption should explain the whole image, including all the main objects, activities, and their relationships. The objects should be named as specifically as practical: For example when describing a young boy in a picture, "young boy" is preferred over "young child", which in turn is preferred over "person".

Note: the goal is to generate captions that would be labeled as "Excellent" under the Rating guidelines above, but raters should not copy captions from the first phase. We want the raters to generate the captions on their own.

We outline here a procedure that you should try and follow when writing your image caption. Note that not all these steps may be applicable for all images, but they should give you a pretty good idea of how to organize your caption. We will make use of the first image in the table below (the one with the young girl smiling) Note: It is acceptable to make assumptions that are reasonable as long as they don't contradict the information in the image (eg: in the second image below, we use "families" in captions 1 and 3 because there seems to be a mix of children and adults though it is not perfectly clear. So it is a reasonable assumption to make and nothing in the image contradicts it. However it is also ok to use "people".)

1. Identify the most salient objects(s)/person(s) in the image; use the most informative level to refer to something (i.e., "girl" rather than "child" or "person"); in the example image: "girl"

Source: 20150726_120207 by Nguyen Hung Vu.

Language Name	Language ID	Caption 1	Caption 2
Arabic	ar	ورود بيجونيا حمراء في الحديقة	صورة مقربة لزهرة بيجونيا حمراء
Bengali	bn	অস্পষ্ট পটভূমিতে গোলাপী রঙের ফুলের কাছের দৃশ্য	
Czech	cs	červeně kvetoucí rostlina	detailní záběr na červené rozkvetlé květiny se zelenými listy
Danish	da	Nærbillede en pink blomst på en plante med grønne og røde blade	nærbillede af begonia
German	de	Nahaufnahme von rosa Begonien und Dach im Hintergrund.	Nahaufnahme von rosa Begonieblumen im Freien, tagsüber.
Greek	el	μπιγκόνια σε κήπο	Κοντινό πλάνο σε άνθος μπιγκόνιας δίπλα και πάνω απο κεραμοσκεπή
English	en	a macro shot of pink flower in the garden	a closeup view of the pink flower with green leaves in garden
Spanish	es	Acercamiento a begonias y hojas de planta con fondo desenfocado.	Acercamiento de una begonia en una maceta en un balcón
Farsi	fa	دو گل بگونیا در زمینه ی با غ بگونیا و زمین مات اطرافش	گل‌های شکفته‌ی قرمز بگونیا روی بوته در باغچه
Finnish	fi	punaisia kukkia terassilla	Lähikuva pinkistä begonia kasvista kasvamassa kukkapenkissä
Filipino	fil	bulaklak na kulay rosas na nasa hardin o paso	Maliit na bulaklak na kulay rosas na may pangalan na begonia ang nakatanim
French	fr	Des fleurs de bégonia dans un jardin	fleurs rose en gros plan avec en arrière plan plusieurs autres flou
Hindi	hi	गुलाबी रंग के फूल का करीबी दृश्य है और पृष्टभूमि में पौधे का धुंधला नज़ारा है	गुलाबी फूलो का करीबी नज़ारा
Croatian	hr	nekoliko cvjetića crvene boje na biljci lončanici	crveni cvijet na balkonu
Hungarian	hu	Kép elmosódott begóniáról a kertben.	piros begónia
Indonesian	id	Bunga berwarna merah muda yang sedang mekar di antara bunga bunga lainnya dengan latar belakang buram	tanaman bunga begonia merah mekar di dalam pot dengan latar belakang atap rumah
Italian	it	fiori rosa con buobo giallo e foglie con sfumature di colore autunnale	fiore rosa con pistillo giallo e foglie verdi in giardino
Hebrew	iw	פרח ורוד וברקע פרחים זהים ועלים ירוקים לא בפוקוס.	פרח בגוניה פורח
Japanese	ja	赤いベゴニアの花のアップ	ベゴニアの花のクローズアップ
Korean	ko	분홍색 꽃이 열린 나무	붉은 색 베고니아꽃 근접샷
Maori	mi		
Dutch	nl	Een roze bloem	Roze begoniabloemen
Norwegian	no	begonia blomster med grønne blader	Begonia blomst
Polish	pl	Czerwone małe kwiaty w ogrodzie	Zbliżenie na przepiękną różową begonię woskową.
Portuguese	pt	flor rosa em destaque, mais flores iguais e folhas verdes desfocadas ao fundo	Jardim com flores vermelhas tendo uma delas em evidência.
Quechua	qu	puka rusadu t´ika wiñashan	Begonia waytakuna jardinpi
Romanian	ro	begonii roz înflorite în prim-plan și pe fundal	begonii roz in gradina
Russian	ru	цветы розовой бегонии	Розовая бегония, цветущая в саду, на фоне кустов других цветов
Swedish	sv	begonia i trädgården	Begoniaväxt.
Swahili	sw	Maua mekundu ya begonia yaliyo pandwa juu ya paa ya nyumba	ua ya rangi ya waridi na kuna mimea yenye matawi za kijani na maua mengine za rangi ya waridi nyuma
Telugu	te	ఎర్రని బిగోనియ పుష్పాలు యొక్క చిత్రం.	మొక్కకి పూసిన గులాబీ రంగు బిగోనియా పువ్వుల గల చిత్రం
Thai	th	ภาพถ่ายระยะใกล้ของดอกไม้สีชมพูที่อยู่บนต้น โดยมีฉากหลังเป็นหลังคาอาคาร	ดอกดาดตะกั่วสีแดงบนต้น
Turkish	tr	Yeşil dallarda pembe çiçekler	behaçede bulunan begonya çiçeği yakın çekimi
Ukrainian	uk	яскрава рожева квітка на кущі крупним планом	Квітка бегонія крупним планом в саду
Vietnamese	vi	ảnh chụp cận cảnh những bông hoa thu hải đường màu hồng nở đang nở trên cành trong vườn	cận cảnh bông hoa cánh hồng nhụy vàng với lá xanh viền đỏ phía sau nền mờ hoa và lá vào buổi sáng
Chinese-Simplified	zh	花丛中盛开着的一种红色花瓣、黄色花蕊的植物特写	房屋边花园里的一大片海棠，其中两朵的近景

Figure 5: Example captions in the 36 languages covered in Crossmodal-3600

2. Identify the most salient relation between the main objects; example “girl standing in front of the whiteboard”
3. Identify the main activity depicted; in the example image: “smiling” as an activity (note that this can also be an attribute of the girl), or “standing” as an activity
4. Identify the most salient attributes of the main object(s)/person(s)/activity(es); in the example image: “smiling” and “young” as attributes for the girl
5. Identify the background/context/environment in which the scene is placed; in the example image: “classroom”
6. Put everything together from steps 1-5 above; for the example image: “a smiling girl standing in a classroom”, or “a young girl smiling in a classroom”.