

# Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, Radu Soricut

Google Research

{asht,jponttuset,chillxichen,rsoricut}@google.com

## Abstract

Research in massively multilingual image captioning has been severely hampered by a lack of high-quality evaluation datasets. In this paper we present the Crossmodal-3600 dataset (XM3600 in short), a geographically-diverse set of 3600 images annotated with human-generated reference captions in 36 languages. The images were selected from across the world, covering regions where the 36 languages are spoken, and annotated 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 strong correlation results with human evaluations when using XM3600 as golden references for automatic metrics.

## 1 Introduction

Image captioning is the task of automatically generating a fluent natural language description for 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., 2015; 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 (Barraut et al., 2018), but they are limited to only a few languages that cover a small fraction of the world’s population, while featuring images that severely under-represent the richness and diversity 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 large potential audience around the world.

Creating large training and evaluation datasets in multiple languages is a resource-intensive endeavor. 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). This work also shows that the effectiveness of some of the most reliable automatic metrics for image captioning, such as CIDEr<sup>1</sup> (Vedantam et al., 2015) is severely diminished when applied to translated evaluation sets, resulting in 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 and expensive human evaluations. However, such evaluations cannot usually be replicated across different research efforts, and therefore do not offer a fast and robust mechanism for model hill-climbing and comparison of multiple lines of research.

The proposed XM3600 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 human caption annotations, we have devised a protocol that allows annotators for a specific target language to produce image captions in a style that is consistent across languages; this protocol results in image-caption annotations that are free of direct translation artefacts, an issue that has plagued Machine Translation research for many years and 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 XM3600 bench-

<sup>1</sup>"measures the similarity of a generated sentence against a set of ground truth sentences written by humans." quoted from (Vedantam et al., 2015).



Source: Porsche Museum, Stuttgart by Brian Solis.

English	<ul style="list-style-type: none"> <li>• A vintage sports car in a showroom with many other vintage sports cars</li> <li>• The branded classic cars in a row at display</li> </ul>
Spanish	<ul style="list-style-type: none"> <li>• Automóvil clásico deportivo en exhibición de automóviles de galería <i>(Classic sports car in gallery car display)</i></li> <li>• Coche pequeño de carreras color plateado con el número 42 en una exhibición de coches <i>(Small silver racing car with the number 42 at a car show)</i></li> </ul>
Thai	<ul style="list-style-type: none"> <li>• รถเปิดประทุนหลายสีจอดเรียงกัน ในที่จัดแสดง <i>(Multicolored convertibles line up in the exhibit)</i></li> <li>• รถแข่งวินเทจจอดเรียงกันหลายคันในงานจัดแสดง <i>(Several vintage racing cars line up at the show)</i></li> </ul>

Figure 1: **Sample captions in three different languages** (out of 36 – see full list of captions 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 Spanish “number 42” or the Thai “convertibles” would not be possible when directly translating from the English versions).

mark, we empirically measure its ability to rank image captioning model variations, and show that it provides high levels of agreement with human judgements, therefore validating its usefulness as a benchmark and alleviating the need for human judgement in the future.

Fig. 1 shows a few sample captions for an image in XM3600 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 have released XM3600 under a [CC-BY4.0](https://creativecommons.org/licenses/by/4.0/) license at <https://google.github.io/crossmodal-3600/>.

## 2 The XM3600 Dataset

In this section, we describe the heuristics used for language and image selection, the design of the caption annotation process, caption statistics including quality, and annotator details.

### 2.1 Language Selection

In this section, we describe the heuristic used for selecting the languages. As a first step, we take a quantitative stance and choose 30 languages ( $\mathcal{L}_{30}$ ) roughly based on their percent of web content<sup>2</sup>. As a second step, we consider an additional five languages ( $\mathcal{L}_5$ )<sup>3</sup> to cover low-resource languages with

<sup>2</sup>Arabic (ar), Chinese-Simplified (zh), Croatian (hr), Czech (cs), Danish (da), Dutch (nl), Filipino (fil), Finnish (fi), French (fr), German (de), Greek (el), Hebrew (he), Hindi (hi), Hungarian (hu), Indonesian (id), Italian (it), Japanese (ja), Korean (ko), Norwegian (no), Persian (fa), Polish (pl), Portuguese (pt), Romanian (ro), Russian (ru), Spanish (es), Swedish (sv), Thai (th), Turkish (tr), Ukrainian (uk), Vietnamese (vi).

<sup>3</sup>Bengali (bn), Cusco Quechua (quz), Maori (mi), Swahili (sw), Telugu (te).

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 to the resulting union of languages plus English, for a total of 36 languages.

### 2.2 Image Selection

In this section, we consider the heuristics used for selecting a geographically diverse set of images. For each of the 36 languages, 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 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 succeeds in providing our target number of 100 images from an appropriate region for most of the 36 languages except for Persian (where 14 continent-level images are used) and

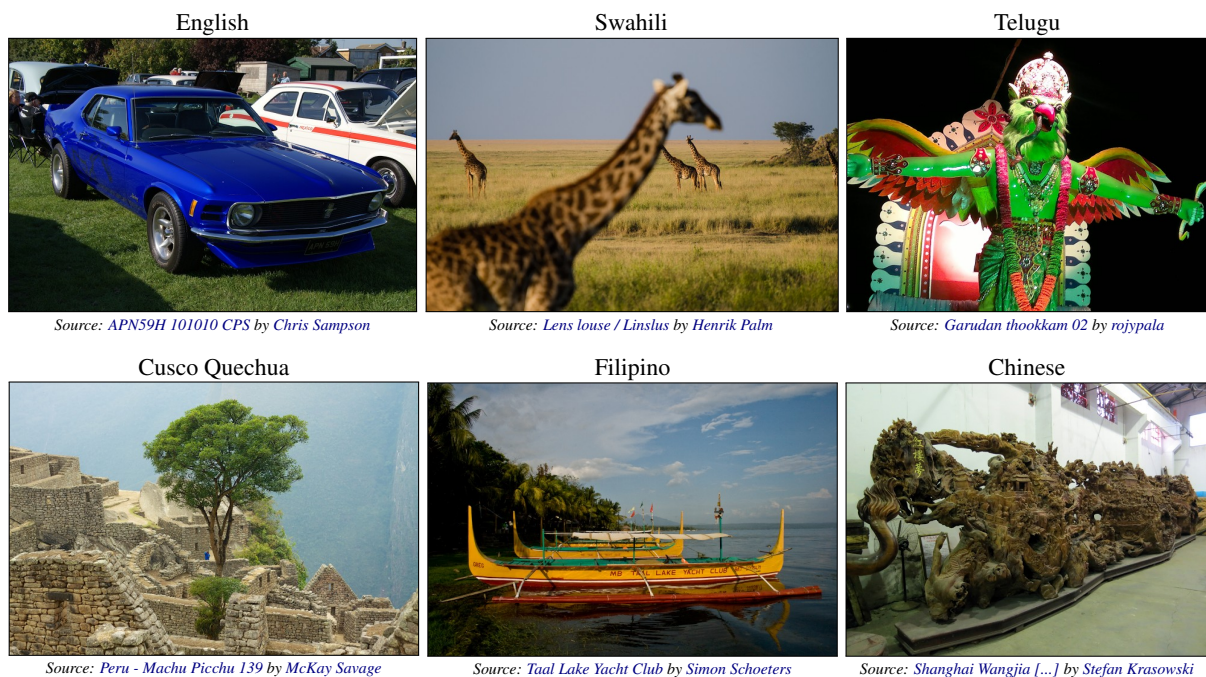


Figure 2: A sample of images in the XM3600 dataset, together with the language for which they have been selected. Overall, the images span regions over 36 different languages and 6 different continents.

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 either evaluate on images relevant to particular regions of interest or on the entire dataset.

### 2.3 Caption Annotation

In this section we detail the design of the caption annotation process. For a massively multilingual benchmark such as XM3600, 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, as it identifies coherence relations between image and captions such as **VISIBLE**, **META**, **SUBJECTIVE**, and **STORY**. The goal for our caption annotation is to generate **VISIBLE** image captions, i.e., use the target language to formulate a sentence that is intended to recognizably characterize what is visually depicted in the image.

One possible approach to generating such captions 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., 2015).

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 refer to this rating scale as the 5-level quality scale in the subsequent text. 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 round shows the same  $N$  images again, but one image at a time 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 the 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  $N$  annotations 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 Spanish mentioning “number 42”, and for Thai mentioning “convertibles”.

We also provide the annotators with an annotation protocol to use when creating the captions, which provides useful guidance in achieving consistent annotations across all the targeted languages. We provide the annotation guidelines in Appendices B and C. For each language, we annotate all 3600 images with captions using replication 2 (two different annotators working independently)<sup>4</sup>, except Bengali (bn) with replication 1 and Maori (mi) with roughly 1 for 2/3 and 2 for 1/3 of the images, see Table 1.

## 2.4 Caption Statistics

In this section, we take a look at the the basic statistics of the captions in the dataset. Table 1 provides detailed caption statistics, including the number of captions per image and the average number of words and characters per caption. There are a total of 261,375 captions across 36 languages, each image having in the vast majority of cases at least 2

<sup>4</sup>Due to various issues related to process idiosyncrasies, the exact replication varies slightly under or over 2.

Lan. Id.	Num. Cap.	Replication			Num. Words	Num. Chars
		1	2	3+		
ar	7367	0	3434	166	7.7	42.2
bn	3600	3600	0	0	11.3	62.1
cs	7207	15	3573	12	6.5	39.1
da	7264	0	3542	58	8.7	48.3
de	8643	0	2240	1360	11.2	76.5
el	7204	0	3596	4	7.7	51.4
en	7200	0	3600	0	9.4	49.5
es	8614	0	2201	1399	9.8	56.3
fa	7245	0	3555	45	12.7	59.4
fi	7127	90	3500	10	7.5	65.2
fil	7109	91	3509	0	12.2	67.6
fr	8562	0	2253	1347	12.3	69.6
he	7200	0	3600	0	11.9	63.6
hi	8503	0	2297	1303	13.4	59.9
hr	7280	0	3553	47	9.0	57.8
hu	7216	0	3586	14	8.5	60.5
id	7126	74	3526	0	14.3	93.5
it	8471	0	2329	1271	12.1	71.8
ja	7185	15	3585	0	1.0	26.0
ko	7650	15	3315	270	7.0	24.7
mi	4732	2483	1102	15	11.7	55.5
nl	8059	0	2771	829	8.0	45.9
no	7213	0	3591	9	9.6	54.3
pl	7141	59	3541	0	8.3	57.6
pt	7243	0	3562	38	10.8	61.7
quz	7200	0	3600	0	5.0	38.6
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	63.0
te	7200	0	3600	0	7.1	47.4
th	7200	0	3600	0	1.2	47.9
tr	7233	15	3538	47	9.4	63.4
uk	7215	0	3585	15	10.0	65.7
vi	7350	0	3450	150	18.0	79.3
zh	7174	60	3508	32	1.0	23.0

Table 1: **Caption statistics:** A total of 261,375 captions across 36 languages. We provide the replication stats per language, as well as average number of words (where applicable) and characters.

captions per language.

For languages with natural space tokenization, the number of words per caption can be as low as 5 or 6 for some agglutinative languages like Cusco Quechua (quz) and Czech (cs), and as high as 18 for an analytic language like Vietnamese (vi). The number of characters per caption 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

In this section, we describe the process for ensuring the creation of high quality annotations, and present

Language	Id	%GOOD+	%MED+	%BAD
Arabic	ar	97.5	99.3	0.7
Bengali	bn	100.0	100.0	0.0
Czech	cs	96.8	99.0	1.0
Danish	da	94.0	99.2	0.8
German	de	98.2	99.3	0.7
Greek	el	77.3	96.0	3.7
English	en	96.5	100.0	0.0
Spanish	es	97.0	98.3	1.7
Farsi	fa	94.0	99.3	0.7
Finnish	fi	91.5	98.8	1.2
Filipino	fil	79.7	95.3	4.5
French	fr	92.7	99.2	0.8
Hebrew	he	82.7	96.7	3.0
Hindi	hi	92.7	98.7	1.3
Croatian	hr	80.7	98.2	1.8
Hungarian	hu	91.3	94.8	5.0
Indonesian	id	90.7	98.5	1.5
Italian	it	88.8	97.7	2.3
Japanese	ja	84.3	96.3	3.5
Korean	ko	85.2	99.5	0.3
Maori	mi	93.5	98.8	1.2
Dutch	nl	92.8	98.7	1.3
Norwegian	no	87.7	96.7	3.3
Polish	pl	92.2	97.3	2.7
Portuguese	pt	87.8	99.5	0.3
Cusco Quechua	quz	83.8	98.3	1.7
Romanian	ro	90.2	98.3	1.7
Russian	ru	93.8	99.5	0.3
Swedish	sv	92.0	99.2	0.8
Swahili	sw	70.0	98.7	1.3
Telugu	te	98.7	99.8	0.2
Thai	th	95.2	99.2	0.8
Turkish	tr	97.8	98.0	1.2
Ukrainian	uk	91.2	99.2	0.8
Vietnamese	vi	94.3	97.8	2.0
Chinese-Simpl.	zh	90.2	97.8	2.2

Table 2: **Caption quality statistics** for the 36 languages. We use the median of three ratings as the aggregated rating for an image-caption pair.

quality statistics of the annotations produced.

In order to ensure quality, the annotation process is initially started with pilot runs on 150 images. The caption ratings are spot checked by the authors to verify that the raters have a good understanding of the rating scale. Further, the generated captions go through a verification round where they are rated by the human annotators on the 5-level quality scale described in Sec.2.3. If the annotations are below the desired quality, we clarify the guidelines and add more examples to provide feedback to the human annotators and then conduct another pilot. This process is repeated until very few low-quality captions are being produced<sup>5</sup>. After this, for every

<sup>5</sup>We started the process with a set of six languages and 4-5 pilots were needed per language. For subsequent languages, only 1-2 pilots were needed because of these clarifications and

language, 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 5-level quality scale mentioned in Sec. 2.3. The quality scores are presented in Table 2.

## 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 given in English.

## 3 Model Comparison using XM3600

In this section, we detail our experiments for comparing several models using human evaluations, and also using XM3600 annotations as gold<sup>6</sup> references for automated metrics.

For model comparison, we train several multilingual image captioning models with different sizes over different datasets, and compare them on XM3600. As our main result, we show a high level of correlation between model rankings based on human-evaluation scores and the scores obtained using CIDEr (Vedantam et al., 2015) with XM3600 annotations as gold references.

### 3.1 Datasets

We build two multilingual datasets for training, CC3M-35L and COCO-35L, by translating Conceptual Captions 3M (Sharma et al., 2018) and COCO Captions (Chen et al., 2015) to the other 34 languages using Google’s machine translation API<sup>7</sup>. The remaining language, Cusco Quechua (quz), is not supported by the API<sup>8</sup>. We use the standard train and validation splits for CC3M<sup>9</sup>. For COCO, we use the *Karpathy split* (Karpathy and Fei-Fei, 2014)<sup>10</sup>.

examples added to the guidelines.

<sup>6</sup>We use “gold” here to refer to human-level quality; references created via means like automatic translation are referred to as “silver” quality.

<sup>7</sup><https://cloud.google.com/translate>

<sup>8</sup>Although the translate.google.com website has recently added support, the API does not support it as of this writing.

<sup>9</sup>CC3M-train: 3,318,333 image-caption pairs. CC3M-validation: 15,840 image-caption pairs.

<sup>10</sup>Train: 113,287 images. Validation: 5000 images. Each image has 5 reference captions.

Model Name	Details	Parameters
BB+CC	mT5-base + ViT-B/16 model pretrained on CC3M-35L and finetuned on COCO-35L	$lr=3e-4, cp=10k$
BB	mT5-base + ViT-B/16 model trained on COCO-35L	$lr=1e-4, cp=10k$
Bg	mT5-base + ViT-g/14 model trained on COCO-35L	$lr=1e-4, cp=10k$
Lg	mT5-large + ViT-g/14 model trained on COCO-35L	$lr=1e-4, cp=10k$

Table 3: **Model details** for all model variants used in our experiments:  $lr$  denotes the learning rate;  $cp$  denotes the number of steps in the constant period where the learning rate is constant.

Model	Lang.	CIDEr XM3600	CIDEr COCO-DEV
BB+CC	en	0.584	0.980
BB	en	0.297	0.856
Bg	en	0.337	0.851
Lg	en	0.343	0.875
BB+CC	es	0.425	0.962
BB	es	0.194	0.844
Bg	es	0.232	0.835
Lg	es	0.220	0.859
BB+CC	hi	0.197	0.759
BB	hi	0.098	0.671
Bg	hi	0.112	0.718
Lg	hi	0.111	0.624
BB+CC	zh	0.202	0.748
BB	zh	0.087	0.659
Bg	zh	0.110	0.695
Lg	zh	0.099	0.656

Table 4: **CIDEr on XM3600 and COCO-DEV** for the models over the four languages  $\mathcal{L}_{CORE}$  (COCO-DEV computed using machine-translated references). Tables 8-11 in the appendix show all the CIDEr values for all the models.

### 3.2 Models

In this section we detail the model architecture we used for the experiments.

Our Transformer-based (Vaswani et al., 2017) model architecture for image captioning is shown in Figure 3. On the vision side, each input image is modeled by a Vision Transformer (ViT) (Dosovitskiy et al., 2020; Zhai et al., 2021). The visual features produced by ViT for every patch of the

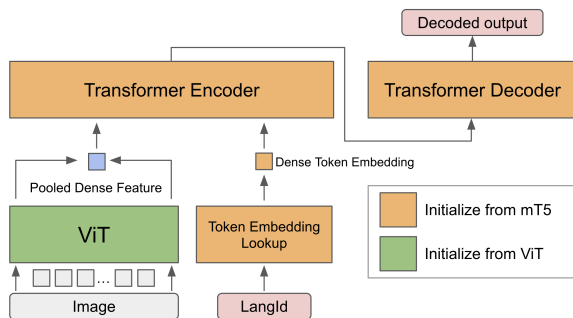


Figure 3: The architecture for the family of multilingual image captioning models used in the experiments.

image are pooled into a single dense feature vector. On the text side, a Language Identifier (LangId) string is used to specify the language. The LangId string is tokenized and embedded into dense token embeddings, which are merged with the dense visual embeddings as the input to a multi-layer Transformer Image and Text Encoder, followed by a multi-layer Transformer Image and Text Decoder to generate the predicted captions.

We take advantage of existing pretrained models to initialize different parts of our model: ViT (Zhai et al., 2021) (green in Fig. 3) and mT5 (Xue et al., 2021) (orange in Fig. 3). We consider different model sizes: mT5-base, mT-large, ViT-B/16, and ViT-g/14, where 16 and 14 are the corresponding patch sizes. We choose three combinations resulting in three different model architectures: mT5-base + ViT-B/16, mT5-base + ViT-g/14 and mT5-large + ViT-g/14.

We train these three models on COCO-35L. In addition, we consider a fourth model based on mT5-base + ViT-B/16 and trained on CC3M-35L. The models are trained on a 4x4x4 TPU-v4 architecture using an Adafactor (Shazeer and Stern, 2018) optimizer with a constant learning rate period between {1k, 10k} steps, followed by a reversed square-root decay with the number of steps. The batch size is 2048 in all the experiments. The initial learning rate is between {1e-4, 3e-4}. We use the same vocabulary (size 250k) as mT5 (Xue et al., 2021). The model trained with CC3M-35L is subsequently finetuned on COCO-35L with constant learning rate 3e-5 for 1 epoch.

Table 3 describes the best hyperparameters we found for different training setups and used in our quantitative experiments. In terms of model sizes, mT5-base has about 680 million parameters, mT5-large about 1230 million, the ViT-B/16 86 million, and ViT-g/14 1011 million parameters. Together, all the experiments took around 5000 TPU hours to train.

### 3.3 Human Evaluation

In this section, we detail the process used for human evaluations comparing the performance of two models.

Our main goal in creating XM3600 is to automate the evaluation of massively multilingual image captioning models, by eliminating expensive and time-consuming human evaluations. Our results indicate that they can be substituted by using the XM3600 annotations as gold references for automated metrics such as CIDEr (Vedantam et al., 2015). To quantify the correlation between the two methods, we train four different models (Tab. 3) and conduct side-by-side human evaluations using the outputs of these models in several languages. We observe strong correlations (Sec. 3.4) between the human evaluations and the CIDEr scores using the XM3600 references.

Specifically, we use a randomly selected subset of 600 images from XM3600 for human evaluations, which we call XM600. Image captions generated by a given pairing of models ( $m_1$  vs  $m_2$ , where  $m_1$  is considered as the base condition and  $m_2$  as the test condition) are compared and rated side-by-side, using a similar pool of annotators as described in Sec. 2.6. Each side-by-side pair (shown in a random per-example left-vs-right order) is rated using a 7-point scale: MUCH-BETTER, BETTER, SLIGHTLY-BETTER, SIMILAR, SLIGHTLY-WORSE, WORSE, MUCH-WORSE, with a replication factor of 3 (three annotators rate each pair). We denote by WINS the percentage of images where the majority of raters (i.e. 2 out of 3) mark  $m_2$ 's captions as better, and by LOSSES the percentage of images where the majority of raters mark  $m_2$ 's captions as worse. We then define the overall side-by-side gain of  $m_2$  over  $m_1$  as  $\Delta S \times S = \text{WINS} - \text{LOSSES}$ .

Conducting the full set of six side-by-side evaluations for each pair of models over the 35 languages would require 210 human evaluation sessions. This is prohibitively expensive and time consuming. Thus, we conduct the full set of six side-by-side evaluations of the pairs of models, on a core set of four languages called  $\mathcal{L}_{\text{CORE}}$ <sup>11</sup>. We call this set of 24 evaluation sessions  $\mathcal{O}_{\text{CORE}}$ . Furthermore, we also conduct a sparser set of side-by-side evaluations over languages where the CIDEr differences on XM3600 and on COCO-DEV<sup>12</sup> indicate

<sup>11</sup>Chinese-Simplified (zh), English (en), Hindi (hi), Spanish (es)

<sup>12</sup>COCO validation split with machine-translated references

disagreement or ambiguity (e.g., opposite sign of the CIDEr differences, and/or small CIDEr differences); this gives us a set of 28 languages called  $\mathcal{L}_{\text{EXT}}$ <sup>13</sup>. We call the resulting set of 41 evaluation sessions  $\mathcal{O}_{\text{EXT}}$ . The set of all evaluations is called  $\mathcal{O}_{\text{ALL}} = \mathcal{O}_{\text{CORE}} + \mathcal{O}_{\text{EXT}}$ , which are conducted over the languages  $\mathcal{L}_{\text{ALL}} = \mathcal{L}_{\text{CORE}} + \mathcal{L}_{\text{EXT}}$ .

The choice of which model is called  $m_1$  and which model is called  $m_2$  is arbitrary in the side-by-side evaluations, since we randomly flip left vs right before presenting the captions to the raters. Hence a single side-by-side evaluation gives two points for the correlation calculations: one with the  $m_1$  and  $m_2$  assigned as per the actual evaluation conducted, and one more with the  $m_1$  and  $m_2$  assignment flipped and the  $\Delta S \times S$  sign flipped correspondingly.

$m_2$	$m_1$	L.	$\Delta S \times S$	$\Delta \text{CIDEr}_{\text{XM600}}$	$\Delta \text{CIDEr}_{\text{XM3600}}$	$\Delta \text{CIDEr}_{\text{COCO-DEV}}$
BB+CC	BB	en	-38.9	-0.277	-0.287	-0.124
BB+CC	Bg	en	-21.5	-0.230	-0.247	-0.129
BB+CC	Lg	en	-34.8	-0.246	-0.240	-0.105
Bg	Lg	en	-3.9	-0.016	0.007	0.024
Bg	BB	en	-14.0	-0.047	-0.039	0.005
Lg	BB	en	-10.3	-0.031	-0.046	-0.018
Bg	Lg	es	-1.3	0.002	-0.012	0.024
Bg	BB	es	-8.4	-0.044	-0.037	0.008
Lg	BB	es	-4.4	-0.045	-0.026	-0.016
BB+CC	Lg	es	-29.6	-0.201	-0.205	-0.103
BB+CC	Bg	es	-28.8	-0.203	-0.193	-0.127
BB+CC	BB	es	-36.5	-0.246	-0.231	-0.118
Bg	Lg	hi	-3.0	-0.001	-0.001	-0.094
BB+CC	Bg	hi	-29.3	-0.095	-0.084	-0.040
Lg	BB	hi	-2.0	-0.012	-0.013	0.047
BB+CC	BB	hi	-36.5	-0.108	-0.099	-0.088
Bg	BB	hi	-5.2	-0.013	-0.015	-0.047
BB+CC	Lg	hi	-32.3	-0.096	-0.086	-0.135
Bg	Lg	zh	-8.9	-0.018	-0.012	-0.039
BB+CC	Lg	zh	-30.0	-0.104	-0.103	-0.092
BB+CC	Bg	zh	-30.1	-0.086	-0.092	-0.053
BB+CC	BB	zh	-41.2	-0.102	-0.115	-0.089
Bg	BB	zh	-16.0	-0.016	-0.023	-0.036
Lg	BB	zh	-11.8	0.002	-0.011	0.003

Table 5: **Model comparisons over  $\mathcal{L}_{\text{CORE}}$  languages** ( $m_2$  vs  $m_1$ ).  $L$  denotes the target language;  $\Delta \text{CIDEr}_{\text{XM600}}$  is  $\text{CIDEr}(m_2) - \text{CIDEr}(m_1)$  on the XM600 dataset,  $\Delta \text{CIDEr}_{\text{XM3600}}$  on the XM3600 dataset, and  $\Delta \text{CIDEr}_{\text{COCO-DEV}}$  on the COCO validation split with machine-translated references. Table 7 in the appendix shows model comparisons over the  $\mathcal{L}_{\text{EXT}}$  languages.

<sup>13</sup>Arabic (ar), Bengali (bn), Croatian (hr), Czech (cs), Danish (da), Dutch (nl), Filipino (fil), Finnish (fi), French (fr), German (de), Greek (el), Hebrew (he), Hungarian (hu), Indonesian (id), Italian (it), Japanese (ja), Korean (ko), Norwegian (no), Persian (fa), Polish (pl), Portuguese (pt), Romanian (ro), Swahili (sw), Swedish (sv), Telugu (te), Thai (th), Turkish (tr), Vietnamese (vi)

Correlation Coefficient	Lang.	N	$\Delta$ CIDEr XM600	$\Delta$ CIDEr XM3600	$\Delta$ CIDEr COCO-DEV
Pearson	$\mathcal{L}_{ALL}$	130	0.88	0.88	0.68
Spearman	$\mathcal{L}_{ALL}$	130	0.87	0.92	0.30
Kendall	$\mathcal{L}_{ALL}$	130	0.69	0.76	0.21
Pearson	$\mathcal{L}_{EXT}$	82	0.72	0.84	-0.44
Spearman	$\mathcal{L}_{EXT}$	82	0.76	0.84	-0.52
Kendall	$\mathcal{L}_{EXT}$	82	0.54	0.65	-0.32
Pearson	$\mathcal{L}_{CORE}$	48	0.90	0.90	0.89
Spearman	$\mathcal{L}_{CORE}$	48	0.95	0.96	0.86
Kendall	$\mathcal{L}_{CORE}$	48	0.80	0.81	0.67

Table 6: **Correlations** between side-by-side human evaluations ( $\Delta S \times S$ ) and CIDEr difference on XM600, XM3600 and the translated COCO validation set. Here N represents the number of points used to compute the correlation coefficient. As noted in Sec. 3.3, each evaluation gives us two points for the correlation calculation.

### 3.4 Results

We present results that show that it is feasible to use the XM3600 annotations as gold references with automated metrics such as CIDEr to compare models in lieu of human evaluations, and that this option is superior to using silver references created via automated translation.

Table 5 presents the results for the  $\mathcal{O}_{CORE}$  set of evaluations on XM600 on the  $\mathcal{L}_{CORE}$  languages, while Table 7 in the appendix shows the results on the  $\mathcal{L}_{EXT}$  languages. The reference for the relative strength of each pairing is given by  $\Delta S \times S$ , with positive numbers indicating the superiority of  $m_2$ , and negative numbers indicating a superiority of  $m_1$ . As can be seen from the table, the model comparisons span a range of model differences, from low  $\Delta S \times S$  to high  $\Delta S \times S$ .  $\Delta$ CIDEr XM600 and  $\Delta$ CIDEr XM3600 capture similar information, except these numbers are based on CIDEr scores using as references XM600 and XM3600, respectively, while  $\Delta$ CIDEr COCO-DEV is based on machine-translated references from the validation split of COCO.

We use the results from Table 5 (and Table 7) to compute the correlation between human judgements of the relative quality of the captioning models and the ability of the CIDEr<sup>14</sup> metric – or, rather, of the underlying references used by the metric – to perform an equivalent task. Table 6 presents the correlation results using three correlation metrics:

<sup>14</sup>We use the reference implementation with default parameters: [github.com/vrama91/cider](https://github.com/vrama91/cider). We remove punctuation and lowercase the captions and references before computing automated metrics.

Pearson, Spearman, and Kendall. The first section shows the correlations over all the side-by-side evaluations (i.e.  $\mathcal{O}_{CORE}$  and  $\mathcal{O}_{EXT}$ ); These cover the  $\mathcal{L}_{CORE}$  and the  $\mathcal{L}_{EXT}$  languages. The second section shows the correlations for the  $\mathcal{O}_{EXT}$  covering the  $\mathcal{L}_{EXT}$  languages. The third section shows the correlations for the  $\mathcal{O}_{CORE}$  evaluations covering the  $\mathcal{L}_{CORE}$  languages.

We observe that  $\Delta$ CIDEr XM3600 is highly correlated with human judgement according to all the correlation metrics (Bonett and Wright, 2000), over all the evaluations  $\mathcal{O}_{ALL}$ , over the  $\mathcal{O}_{CORE}$  evaluations, and also the  $\mathcal{O}_{EXT}$  evaluations. Furthermore, for the  $\mathcal{O}_{EXT}$  evaluations, where most of the instances have opposite signs for  $\Delta$ CIDEr COCO-DEV and  $\Delta$ CIDEr XM3600, we find that the former is strongly anti-correlated with the human evaluation results while the latter is highly correlated with the human evaluation results. Overall, these results indicate that: (i) we can reliably substitute  $\Delta$ CIDEr XM3600 for human evaluations on XM600 when comparing models similar to the ones we used; (ii) the gold XM3600 references are preferable over the silver references obtained from translating COCO captions, in terms of approximating the judgements of the human evaluators<sup>15</sup>.

Based on the results from Table 6, we recommend the use of the XM3600 references as a means to achieve high-quality automatic comparisons between multilingual image captioning models. We have provided the CIDEr scores for XM3600 in 35 languages for all the models, in Tables 8-11 in the Appendix. These can be used as baselines in future work.

<sup>15</sup>However, it is unclear whether machine translated references for one particular language in XM3600 translated to all others, are worse than using the human generated references. In particular, we studied the correlations of CIDEr computed using XM3600-en-MT (i.e. the XM3600 English references, machine translated to all the other languages), with the human evaluations. We found that even though the translations have artifacts and disfluencies, CIDEr differences calculated using them show comparable correlations with human judgement observations. We also studied such correlations for machine translated references from German, Greek, Hebrew, Hungarian and Swahili. We found that the correlations are similar and sometimes even a bit higher than using the human generated references. We believe this happens because the rater guidelines weigh informativeness over fluency and the CIDEr metric is also not as sensitive to fluency. Further work is needed to understand the use of translated references as compared to human generated references. We believe that using the human generated references along with the set of machine translated references from all the other languages may provide even stronger correlations and show greater diversity in the coverage of the image constituents.



## 4 Conclusions

We introduce the XM3600 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 XM3600 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.

Our empirical observations are primarily on the full set of side-by-side comparisons over English and three other languages (Spanish, Hindi, Chinese). Due to the similarity in the data collection and the quality control process, we expect similar results to hold for all the other languages as well; we validated this expectation with additional empirical observations covering an additional 28 languages.

## 5 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. Additionally, since the  $\mathcal{L}_{30}$  languages were selected based on their internet presence, one unintended consequence is that the dataset over-represents European languages. While this is somewhat mitigated by including the  $\mathcal{L}_5$  low resource languages, building and sharing this dataset can have the unintended effect of perpetuating the issue where computational linguistics and AI work is often unintentionally Eurocentric.

Due to the cost and logistical constraints, we have sampled only 100 images for each of the tar-

geted 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. Further, there is the possibility of bias in the dataset due to the uneven access to photographic equipment and internet connectivity (For example, several of the images in Fig. 4 seem to be shared by people with non-native names in the context of the locales. Thus, these images may have been taken by tourists rather than natives. Further exploration into this aspect of the dataset is important as well).

Another limitation is around the absence of translation artifacts in the annotations. We primarily rely on the caption generation process outlined in Sec. 2.3 and on rater quality controls for avoiding translation artifacts. Further, we have performed spot checks on captions in several languages and have not found indications of translation artifacts. Additionally, we have also compared the translations of annotations from another language such as English with generated annotations and verified that the translations show peculiar artifacts and disfluencies which are not seen in the generated annotations.

We would also like to emphasize that, while this dataset aims to ameliorate the need for human evaluations for multilingual image captioning, automated evaluation may be less sensitive to small changes, e.g. when comparing highly tuned methods submitted to competitions. This was one of our motivations for comparing models that range from very different (CC+B<sub>g</sub> vs B<sub>g</sub>/L<sub>g</sub>/BB) to moderately different (BB vs B<sub>g</sub>/L<sub>g</sub>) and quite similar (B<sub>g</sub> vs L<sub>g</sub>), and the results from Table 5 show that our approach works well over this range of model differences over  $\mathcal{L}_{CORE}$ . We also stress-tested our approach by focusing the  $\mathcal{O}_{EXT}$  evaluations on cases where  $\Delta CIDEr_{XM3600}$  or  $\Delta CIDEr_{COCO-DEV}$  were quite small or of opposite signs, and the results from Table 7 in the appendix show that  $\Delta CIDEr_{XM3600}$  correlated well with human evaluations even for this harder set of evaluations. However, we caution the reader that there will be cases where human judgement will still be needed. Further, automated evaluations may be biased to methods that explicitly optimize the evaluated metric, e.g. via approaches such as Self-Critical Sequence Training (Rennie et al., 2017).

We also note that the model outputs and human

judgements data used for calculating the correlations would be useful for constructing new automated metrics and validating existing automated metrics for model comparisons. Releasing this data would also allow independent calculation of CIDEr and  $\Delta S \times S$  shown in Table 5 and Table 7. However, due to the timelines involved and approvals required, we are not able to release this data with the paper. This may hamper the reproducibility of these computations.

The approach to data collection and annotation of COCO-CAP (Chen et al., 2015) and CC3M (Sharma et al., 2018) upholds rigorous privacy and ethics standards, such as the avoidance of offensive content and exposure of personal identification data. This significantly mitigates but does not completely eliminate the risks that the captioning models we train would produce such information. Similarly, the XM3600 dataset mitigates such risks by adopting a defense-in-depth approach: 1) The annotations have been produced in-house and have been quality controlled, while the images used have been vetted to be appropriate for the intended use. 2) Further, the machine translations of the annotations have been scanned with an automated tool to detect personally identifiable information. 3) The machine translations of the annotations have been spot-checked by the authors.

Overall, in spite of the above limitations, we believe that this dataset is a significant step toward ameliorating language and geographic bias, and that it should be used for advancing image captioning research over a wider variety of images and languages.

## 6 Acknowledgements

We would like to thank the anonymous reviewers for providing feedback which led to several improvements such as: 1) A discussion about correlations of human judgement with the machine translations of the XM3600 references; 2) A discussion about the possibility of releasing model outputs and human evaluation data which may help with reproducibility and also help evaluation of existing and new automated metrics over  $\mathcal{L}_{ALL}$ .

## 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.

- 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.
- Douglas Bonett and Thomas Wright. 2000. [Sample size requirements for estimating pearson, kendall and spearman correlations](#). *Psychometrika*, 65:23–28.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. [An image is worth 16x16 words: Transformers for image recognition at scale](#).
- 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.
- Andrej Karpathy and Li Fei-Fei. 2014. [Deep visual-semantic alignments for generating image descriptions](#).
- 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 Mallocci, 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*.

- Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives.
- Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. [Self-critical sequence training for image captioning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.
- 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.
- Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*.
- 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. In *NeurIPS*.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- 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.
- Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2021. [Scaling vision transformers](#).
- 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 4 displays the captions in the 36 languages covered in XM3600 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):

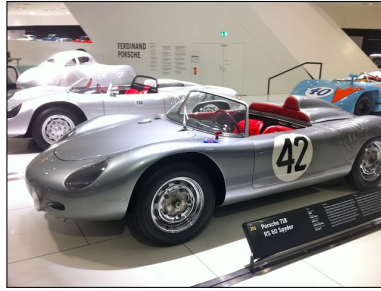
- **BAD:** The caption has one or more of the following issues: a). Caption misses the main topic of the image. b). Caption has major grammatical errors (such as being incomplete, words in wrong order, etc). Please ignore capitalization of words and punctuation. c). Caption violates the ‘No Hallucination’ 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.



Source: Porsche Museum, Stuttgart by Brian Solis.

Language Name	Language ID	Caption 1	Caption 2
Arabic	ar	سيارة رمادي رياضية قديمة في معرض سيارات	سيارة جديدة وحديثة تقف في صف في أحد معارض السيارات وهناك أيضاً سيارات أخرى ملونة معروضة
Bengali	bn	একটি হলের মধ্যে সাদা ও সিলভার রঙের স্পোর্টস কার রাখা আছে	
Chinese-Simplified	zh	在展厅里停靠着一排老爷车正在展出，离得最近的是这一辆灰色的	车展中都是保时捷敞篷跑车
Croatian	hr	sportski automobil sive metalik boje tipa kabriolet izložen u muzeju	trkači auti na izložbi, sivi Porsche s brojem 42
Cusco Quechua	quz	huk uqi karru mana tichuyuq	Ñawpaq phawana carrukuna musuqlana k'anchasqa qhawakushan
Czech	cs	historická závodní auta	retro modely Porsche na výstavě aut
Danish	da	En ældre lavere sølvfarvet racerbil på en udstilling med andre racerbiler	Udstilling med veteranbiler
Dutch	nl	klassieke raceauto's op een rij een museum	Klasieke race auto's op een rij in een showroom
English	en	The branded classic cars in a row at display.	A vintage sports car in a showroom with many other vintage sports cars.
Persian	fa	تعدادی اتومبیل مسابقه ای اسپرت در یک نمایشگاه سرپوشیده	چندین خودروی اسپورت کمپانی پورشه که در کنار هم در یک موزه پارک شده اند
Filipino	fil	mga klasikong sasakyan na nakadisplay sa tindahan ng kotse	magkakatabing mga lumang kotse
Finnish	fi	Antiikkisia urheiluautoja näytellyssä	
French	fr	Une série de voitures de course vintage exposé dans un musée	voiture ancienne de course grise numéroté 42 exposé au côté d'autres voiture en intérieur
German	de	Verschiedenfarbige Rennwagen mit Nummern auf einer Autoausstellung	Ein silber-metallic Cabrio Oldtimer Porsche 718 steht auf einem Flachen Podest mit anderen Oldtimer dahinter im Porsche Museum in Stuttgart
Greek	el	έκθεση ρετρό αγωνιστικών αυτοκινήτων	Κλασικό γκρι αγωνιστικό καμπριολέ αυτοκίνητο σε έκθεση δῖπλα σε άλλα.
Hebrew	he	רכב פורשה וינטג' עם גג נפתח בצבע כסף עומד בתצוגה ליד רכבים נוספים.	מכוניות ספורט חדשניות עומדות בתצוגה
Hindi	hi	सफ़ेद फर्श पर ठेरी कोरे रंग की गाडी और उसके पीछे और भी गाड़ियां	हॉल में लगी नए ज़माने की गाड़ियों का दृश्य .
Hungarian	hu	Veterán sportautók egymás mellett állnak egy kiállítóteremben	Klasszikus sportautók kiállítva egy múzeumban
Indonesian	id	Sebuah pameran mobil konvertibel klasik di mana terdapat deretan mobil yang diparkir di dalam ruangan	deretan beberapa mobil balap konvertibel klasik Porche dipajang di museum mobil klasik
Italian	it	auto sportive d'epoca in esposizione in un salone dalle mura bianche	auto da competizione di qualche decennio fa esposte a salone dell'auto
Japanese	ja	銀色のメタリック・スポーツカーと、他の展示されている車	ポルシェミュージアムに展示されている複数のオープンカー
Korean	ko	내부에 포르쉐 자동차가 진열되어 있다	포르쉐 스포츠카 전시장에 진열된 번호가 달린 다양한 차들
Maori	mi	Etaki motuka tawhito kei roto i tetahi whare	
Norwegian	no	klassiske sportsbiler på rad presentert i et galleri	Lav klassisk sportsbil i sølv ved siden av andre biler i utstillingshall
Polish	pl	Retro samochody sportowe na wystawie	sportowe samochody rajdowe na wystawie
Portuguese	pt	exposição de carros porsche com o modelo 718 a frente	Carros modernos enfileirados em uma exposição
Romanian	ro	mașini de jucărie sport colorate diferite aranjate la expoziție pe raftul alb	masina clasica gri decapotabila porsche cu alte masini similare parcate in spatele ei in interiorul unei camere cu pereti si podea albe
Russian	ru	серебристый спортивный автомобиль марки Порше с красным салоном на выставк	Спортивный автомобиль цвета металллик с открытым верхом в выставочном зале на фоне других, стоящих в ряд, спортивных автомобилей
Spanish	es	Automóvil clásico deportivo en exhibición de automóviles de galería	Coche pequeño de carreras color plateado con el número 42 en una exhibición de coches
Swahili	sw	magari matatu ya klasiki yaliyopangwa kwa mfululizo	Magari wa muundo wa zamani wa kipekee yakiwa yamepangwa mfululizo kwa kando kando yakiwa kwenye maonyesho
Swedish	sv	Flera klassiska sportbilar i rad inomhus	Tre porsche sportmodell utan tak i ljus utställningslokal med fler bilar i bakgrunden
Telugu	te	వరుసలో ఉన్న పిల్లలు ఆడుకునే కార్లు	అంగడిలో ప్రదర్శనలో వరుసగా పెట్టిన నూతన కార్ల యొక్క చిత్రం
Thai	th	รถเปิดประทุนหลายสีจอดเรียงกันในพื้นที่จัดแสดง	รถแข่งเงินเทาจอดเรียงกันหลายคันในงานจัดแสดง
Turkish	tr	Galerideki eski Porsche yarış arabaları	Kapalı alanda duran kırmızı koltuklu gri renkli üzeri açık bir spor araba ve arkasında duran ona benzer başka spor arabalar
Ukrainian	uk	сірий спортивний ретро автомобіль в автосалоні на тлі таких іншого кольору	Спортивні ретро автомобілі Порше в музеї
Vietnamese	vi	những chiếc xe mui trần sang trọng được trưng bày cạnh nhau trong một khu triển lãm	mô hình xe hơi mui trần nhiều màu trên nền gạch trắng phía trước có tấm bảng đen chữ trắng

Figure 4: Example captions in the 36 languages covered in XM3600

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"
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".

## D Detailed Results for Model Comparison

$m_2$	$m_1$	L	$\Delta S \times S$	$\Delta \text{CIDEr}_{\text{XM600}}$	$\Delta \text{CIDEr}_{\text{XM3600}}$	$\Delta \text{CIDEr}_{\text{COCO-DEV}}$
Lg	BB	ar	-1.2	-0.003	-0.003	0.055
Lg	BB	bn	-3.5	-0.025	-0.026	0.039
Bg	BB	cs	-5.2	-0.031	-0.016	0.013
Lg	BB	cs	-2.7	-0.012	0.002	0.029
Bg	BB	da	-6.9	-0.021	-0.029	0.048
Bg	Lg	da	2.7	-0.009	-0.003	0.029
Lg	BB	da	-13.3	-0.012	-0.026	0.018
Bg	BB	de	-12.6	-0.014	-0.026	0.030
Bg	Lg	de	-1.5	-0.007	-0.008	0.037
Lg	BB	de	-9.6	-0.007	-0.018	-0.006
Lg	BB	el	-5.1	-0.005	0.002	0.063
Lg	BB	fa	-11.1	-0.011	-0.003	0.027
Lg	BB	fi	-0.3	-0.007	-0.006	0.008
Lg	BB	fil	-3.2	-0.024	-0.004	0.020
Lg	BB	fr	-2.0	-0.030	-0.015	0.011
Bg	BB	fr	-3.0	-0.022	-0.024	0.001
Lg	BB	he	1.7	0.005	0.001	0.025
Lg	BB	hr	-4.5	-0.007	0.002	0.030
Bg	BB	hr	-8.4	-0.019	-0.023	0.014
Lg	BB	hu	-4.7	-0.005	-0.009	0.027
Lg	BB	id	-4.7	-0.031	-0.018	0.004
Lg	BB	it	-6.2	-0.011	-0.015	0.010
Lg	BB	ja	-10.8	-0.013	-0.021	-0.006
Lg	BB	ko	-2.0	-0.025	-0.018	0.045
Bg	BB	nl	-7.7	0.003	-0.016	0.009
Bg	Lg	nl	0.2	0.001	-0.003	0.007
Lg	BB	nl	-6.2	0.002	-0.013	0.002
Lg	BB	no	-10.3	-0.043	-0.033	0.007
Bg	BB	pl	-3.5	-0.022	-0.023	0.015
Lg	BB	pl	-4.4	0.002	-0.007	0.019
Lg	BB	pt	-10.8	-0.027	-0.026	-0.011
Lg	BB	ro	-5.6	-0.027	-0.017	-0.001
Lg	BB	sv	-8.1	-0.028	-0.035	0.003
Bg	Lg	sv	-5.4	0.004	0.004	0.031
Bg	BB	sv	-9.3	-0.024	-0.032	0.035
Lg	BB	sw	-2.2	-0.016	0.007	0.040
Lg	BB	te	-1.3	0.008	-0.002	0.037
Lg	BB	th	-6.7	-0.031	-0.016	0.028
Lg	BB	tr	-5.1	-0.026	-0.016	0.029
Bg	BB	vi	-4.5	0.000	0.000	0.058
Lg	BB	vi	3.2	0.015	0.008	0.082

Table 7: **Model comparison over the  $\mathcal{L}_{\text{EXT}}$  languages ( $m_2$  vs  $m_1$ ).**  $L$  denotes the target language;  $\Delta \text{CIDEr}_{\text{XM600}}$  is  $\text{CIDEr}(m_2) - \text{CIDEr}(m_1)$  on the XM600 dataset,  $\Delta \text{CIDEr}_{\text{XM3600}}$  on the XM3600 dataset, and  $\Delta \text{CIDEr}_{\text{COCO-DEV}}$  on the COCO validation split with machine-translated references.

Lang.	CIDEr	CIDEr
	XM3600	COCO-DEV
ar	0.227	0.649
bn	0.200	0.682
cs	0.313	0.575
da	0.329	0.877
de	0.224	0.735
el	0.199	0.830
en	0.584	0.980
es	0.425	0.962
fa	0.311	0.898
fi	0.177	0.487
fil	0.353	1.007
fr	0.410	0.957
he	0.230	0.650
hi	0.197	0.759
hr	0.224	0.607
hu	0.175	0.551
id	0.307	1.088
it	0.321	0.902
ja	0.254	0.963
ko	0.288	0.862
mi	0.405	1.175
nl	0.441	0.796
no	0.385	0.856
pl	0.236	0.578
pt	0.380	0.964
ro	0.188	0.832
ru	0.194	0.675
sv	0.370	0.848
sw	0.319	0.796
te	0.196	0.520
th	0.418	0.929
tr	0.232	0.668
uk	0.189	0.653
vi	0.336	1.150
zh	0.202	0.748

Table 8: **CIDEr on XM3600 and COCO-DEV** for the best performing model BB+CC on all 35 languages. (COCO-DEV computed using machine-translated references).

Lang.	CIDEr	CIDEr
	XM3600	COCO-DEV
ar	0.121	0.573
bn	0.139	0.623
cs	0.157	0.500
da	0.195	0.736
de	0.138	0.612
el	0.119	0.777
en	0.337	0.851
es	0.232	0.835
fa	0.180	0.816
fi	0.098	0.442
fil	0.215	0.951
fr	0.226	0.835
he	0.121	0.584
hi	0.112	0.718
hr	0.111	0.509
hu	0.107	0.491
id	0.187	1.000
it	0.184	0.783
ja	0.154	0.900
ko	0.169	0.787
mi	0.261	1.121
nl	0.235	0.697
no	0.242	0.768
pl	0.125	0.499
pt	0.222	0.852
ro	0.105	0.737
ru	0.111	0.588
sv	0.221	0.717
sw	0.191	0.702
te	0.112	0.497
th	0.253	0.856
tr	0.141	0.636
uk	0.091	0.631
vi	0.190	0.964
zh	0.110	0.695

Table 9: **CIDEr on XM3600 and COCO-DEV** for the model Bg on all 35 languages. (COCO-DEV computed using machine-translated references).

Lang.	CIDEr	
	XM3600	COCO-DEV
ar	0.106	0.513
bn	0.133	0.555
cs	0.139	0.485
da	0.192	0.765
de	0.130	0.649
el	0.101	0.680
en	0.343	0.875
es	0.220	0.859
fa	0.155	0.766
fi	0.089	0.419
fil	0.185	0.858
fr	0.217	0.825
he	0.098	0.548
hi	0.111	0.624
hr	0.085	0.493
hu	0.096	0.451
id	0.167	0.943
it	0.168	0.770
ja	0.141	0.850
ko	0.152	0.716
mi	0.243	0.942
nl	0.232	0.704
no	0.230	0.736
pl	0.108	0.495
pt	0.202	0.843
ro	0.100	0.709
ru	0.089	0.581
sv	0.225	0.748
sw	0.151	0.640
te	0.099	0.426
th	0.226	0.802
tr	0.122	0.584
uk	0.081	0.560
vi	0.182	0.940
zh	0.099	0.656

Table 10: **CIDEr on XM3600 and COCO-DEV** for the model Lg on all 35 languages. (COCO-DEV computed using machine-translated references).

Lang.	CIDEr	
	XM3600	COCO-DEV
ar	0.103	0.568
bn	0.107	0.594
cs	0.141	0.514
da	0.166	0.783
de	0.112	0.643
el	0.103	0.743
en	0.297	0.856
es	0.194	0.844
fa	0.152	0.793
fi	0.083	0.427
fil	0.181	0.878
fr	0.202	0.836
he	0.099	0.573
hi	0.098	0.671
hr	0.087	0.523
hu	0.087	0.478
id	0.149	0.947
it	0.153	0.780
ja	0.119	0.844
ko	0.134	0.760
mi	0.239	1.049
nl	0.219	0.705
no	0.197	0.743
pl	0.101	0.515
pt	0.176	0.832
ro	0.084	0.709
ru	0.077	0.586
sv	0.189	0.752
sw	0.158	0.681
te	0.097	0.463
th	0.210	0.830
tr	0.106	0.613
uk	0.081	0.580
vi	0.190	1.022
zh	0.087	0.659

Table 11: **CIDEr on XM3600 and COCO-DEV** for the model BB on all 35 languages. (COCO-DEV computed using machine-translated references).