Denoising Large-Scale Image Captioning from Alt-text Data using Content Selection Models

Khyathi Raghavi Chandu†, Piyush Sharma‡, Soravit Changpinyo*, Ashish Thapliyal*, Radu Soricut*

† Meta AI, ‡ Uber, * Google Research

khyathi.research@gmail.com, (schangpi, asht, rsoricut)@google.com

Abstract

Training large-scale image captioning (IC) models demands access to a rich and diverse set of training examples that are expensive to curate both in terms of time and man-power. Instead, using alt-text based captions gathered from the web is a far cheaper alternative for scaling with the downside being that the data is noisy. Recent modeling approaches to IC often fall short in terms of performance in leveraging these noisy datasets as compared to datasets with clean annotations. We address this problem with a simple yet effective technique of breaking down the task into two smaller, more controllable tasks – skeleton prediction and skeleton-based caption generation. Specifically, we show that sub-selecting content words as skeletons helps in generating improved and denoised captions when leveraging rich yet noisy alt-text-based uncurated datasets. We also show that the predicted English skeletons can further cross-linguistically be leveraged to generate non-English captions, by presenting experimental results in French, Italian, German, Spanish and Hindi. We also show that skeleton-based prediction allows for better control of caption properties, such as length, content, and gender expression, providing a handle to perform human-in-the-loop interpretable semi-automatic corrections.

1 Introduction

In the last demi-decade, NLP fields have ventured into reaping the benefits of utilizing large scale raw (uncurated) data from web-crawls. This trend aligned with new uncurated image-captioning datasets like Conceptual Captions (Sharma et al., 2018). While these uncurated datasets are superior in terms of size and diversity, they are inferior to well curated datasets (Lin et al., 2014; Wang et al., 2019b) in terms of noise in the captions. The

![Image Captioning Pipeline](image.png)

Figure 1: Overview of our approach: (1) skeleton prediction & (2) skeleton based IC; compared to conventional IC. Output captions shown in English (En), Hindi (Hi) and Italian (It).

content in the alt-text for the image is often distorted by the intent or context in which the image is presented. For example, the ground truth alt-text caption for a house is ‘house for sale’ instead of ‘front view of a house’. This hinders the use of these large noisy datasets to the fullest extent.

We present a simple two-staged approach by separating the content selection from caption generation as illustrated in Figure 1. In contrast to most IC approaches (Hossain et al., 2018; Sharma et al., 2020), which hallucinate incorrect content from noisy training data (i.e. ‘custom posters’ and ‘wedding’), our approach first focuses on denoising the content words (i.e ‘collection’ and ‘book’) that are further used to generate a relevant caption. We refer to this sequence of concept words that are key pieces of information consistent with the image as a skeleton. Sub-selecting skeleton words that curb noisiness are automatically extracted from the alt-text captions. We focus on language-based skeletons that are derived from captions (Kuznetsova et al., 2014; Fang et al., 2015; Dai et al., 2018), rather than expensive visual-based skeletons derived from image, e.g. scene graphs, (Wang et al., 2019a; Yang et al., 2019), which are hard to scale. More concretely, we introduce an intermediate task of distantly supervised skeleton prediction in the end to end IC pipeline: The end-to-end task of IC ($f_o : \mathbb{I} \rightarrow \mathbb{C}$) is broken down into a two-staged pipeline: skeleton prediction ($f_{\theta} : \mathbb{I} \rightarrow \mathbb{S}$) and
skeleton based captioning \((f_\phi : \mathbb{I}, \mathbb{S} \rightarrow \mathbb{C})\), where \(\mathbb{I}\) is the image, \(\mathbb{S}\) is the skeleton, and \(\mathbb{C}\) is the caption \((\text{Kulkarni et al., 2013; Li et al., 2011; Elliott and Keller, 2013; Fang et al., 2015})\). We present a comparison between encoding, decoding and autencoding these skeletons. As such, our skeleton prediction solution addresses the semantic gap problem \((\text{Li and Chen, 2018; Yao et al., 2018})\).

We illustrate the effectiveness of this approach on uncurated noisy datasets in the following ways. (1) We demonstrate that sub-selecting content words with an intermediate skeleton prediction task denoises content thereby leading to better human evaluation results on captioning. We also conduct an extensive analysis on multimodal discourse relations and find that the reason for this improvement is the generation of more visible captions \((\text{Alikhani et al., 2020})\). (2) Scaling large uncurated datasets to other languages is still a bottleneck. We show the transferability of learning English skeletons to improve caption generation in other languages – English, French, Italian, German, Spanish and Hindi. (3) The predicted skeletons qualitatively demonstrate other potential benefits, such as controllability of content, length, and gender via a natural language–based interpretable interface, which enables one to additionally interact with the generation process.

2 Related Work

Content selection from vision: There is a rich body of work in improving content selection for IC \((\text{Feng et al., 2019})\), mainly focused on scene graph based skeletons \((\text{Gu et al., 2019; Kim et al., 2019; Chen et al., 2020a; Yang et al., 2019})\). However, these annotations with objects and relations are expensive, thereby constraining the scaling up to multiple languages and diverse concepts. Our work delegates this responsibility of identifying content to the language modality by using inexpensive off the shelf tools for weak supervision.

Content selection from language: An orthogonal body of work relies on skeletons derived from language using hierarchical phrase modeling \((\text{Tan and Chan, 2016; Dai et al., 2018})\), semantic attention \((\text{You et al., 2016})\), attribute LSTM \((\text{Yao et al., 2017})\), skeleton based attribute filling \((\text{Wang et al., 2017})\), adaptively merging topic and visual information \((\text{Liu et al., 2018})\), multimodal flow \((\text{Li et al., 2019a})\) and concept guided attention \((\text{Li et al., 2019b})\). Note that all these prior works utilize human curated gold datasets such as COCO \((\text{Lin et al., 2014})\) and Flickr30k \((\text{Plummer et al., 2015})\) with clean coupling between captions and images. However, scaling them to large and diverse concepts is expensive. We utilize uncurated silver standard datasets with the advantages of richness and diversity at the cost of noisy text. Hence we show the effectiveness of a dual staged approach that denoises the captions by skeleton prediction.

Cross-lingual and controllable captions: Past work on cross-lingual captioning focused on translation \((\text{Barrault et al., 2018})\), fluency guidance \((\text{Lan et al., 2017})\), using large datasets \((\text{Yoshikawa et al., 2017})\) and more recently by pivoting on source language captions \((\text{Thapliyal and Soricut, 2020; Gu et al., 2018})\). We go a step further and pivot on the predicted English skeleton to improve multilingual captions due to the dearth of similar off the shelf tools in other languages. We qualitatively explore controlling length via skeletons which was explored before via adding length to decoder \((\text{Luo and Shakhnarovich, 2020; Cornia et al., 2019})\). Other controllable aspects include stylistic captions \((\text{Guo et al., 2019; Mathews et al., 2018})\) language \((\text{Tsutsui and Crandall, 2017})\) which are potential extensions to our work.

Interpretable Natural language skeletons: Despite remarkable advancements of large scale end-to-end models, recent work identifies spurious correlations in datasets that potentially lead to high performance \((\text{Geva et al., 2019; Tsuchiya, 2018})\). To mitigate this, researchers began to dissect intermediate components of models with the goal of interpretability to humans \((\text{Wiegrefe and Pinter, 2019; Thorne et al., 2019; Lipton, 2018})\) as opposed to implicit explanation \((\text{Xu et al., 2015})\). Our work is an instance of explaining captions through skeleton predictions similar to recent work on rationalizing answer predictions for question answering \((\text{Latcinnik and Berant, 2020})\). We view the intermediate skeleton layer as an interpretable model prediction that helps us study key subtle dataset attributes, such as gender bias.

3 Our Approach

IC requires paired examples of images and captions \((\mathbb{I}, \mathbb{C})\), where \(e \in \mathbb{C}\) correspond to tokens in a caption \((c_1, c_2, \ldots, c_m)\), which are often expensive to gather. Under this paradigm, end-to-end model training attempts to perform a match between the
semantic concepts present in \( I \) and \( C \), starting from image, region, and object level features and mapping them to various \( c_i \)s. In contrast, our approach uses intermediate skeletons as an effective way to leverage noisy, uncurated alt-text based captions to train a model to generate more visually informative captions. An overview of both the stages is presented in Fig. 1.

3.1 Distantly Supervised Skeletons
Since gold standard skeleton words are usually not available, we use distant supervision to get these labels. We retrieve syntax annotations (POS tags and word lemmas), using the Google Cloud Natural Language API\(^1\) of caption texts. We use these annotations to experiment with skeleton variants. The ground-truth skeletons are selected by analyzing the syntax of the automatically curated web-scaled captions through combinations of nouns, verbs, adjectives and adverbs in their condensed forms. In addition, we also ignore tokens with a frequency of less than 50 in our training data to reduce noise while selecting the skeleton words. This subselection of content based on POS tags and downscaling of vocabulary helps in retaining important words as skeletons resulting in a label size of 5k. Since automatic n-gram based metrics cannot be evaluated against noisy ground-truths, manual evaluation is conducted to understand the denoising of sub-selection.

1. **Nouns & Verbs:** This includes a sequence of lemmas of all the nouns and verbs in a caption.
2. **Salient Nouns & Verbs:** Saliency of nouns and verbs is determined using tf-idf scores, treating each caption as a document. For each caption, the top 2 highest scoring noun and verb tokens (lemma) are selected. This examines if saliency contributes towards effectiveness of the skeleton.

3. **Nouns:** This includes lemmas of all the nouns. This helps us untangle the roles of nouns vs verbs in the effectiveness of the skeleton.

4. **Iteratively refined captions:** Under this condition, the output of the baseline Img2Cap model serves as the ‘skeleton’ for the next skeleton-based captioning stage. The rationale behind this skeleton is to compare the utility of sub-selecting skeleton words based on POS in denoising caption content, compared to a full caption prediction.

3.2 Model
**Baseline (Img2Cap):** We adopt an encoder-decoder \( (f_\theta : I \rightarrow C) \) IC model based on Transformers (Vaswani et al., 2017) following recent state-of-the-art approaches (Sharma et al., 2018; Yu et al., 2019; Changpinyo et al., 2019; Huang et al., 2019; Cornia et al., 2020). Our model uses the IC framework introduced in (Changpinyo et al., 2019). Inspired by the bottom-up and top-down approach (Anderson et al., 2018), the input image \( I \) is represented as a bag of features, containing one global and 16 regional, fine-grained feature vectors. The regional features correspond to the top 16 box proposals from a Faster-RCNN (Ren et al., 2015) object detector trained on Visual Genome (Krishna et al., 2017), with a ResNet101 (He et al., 2016) that is trained on JFT (Hinton et al., 2015) and fine-tuned on ImageNet (Russakovsky et al., 2015). We featurize both global and regional boxes using Graph-RISE (Juan et al., 2019, 2020). We

\(^1\)https://cloud.google.com/natural-language
An attractive property is that the same architecture with a Linear-ReLU-LayerNorm-Linear instead of the second stage of training uses images and skeletons of the skeleton words (from §3.1) from $\mathcal{S}$ to generate captions $\hat{C}$. The second stage of training uses images and skeletons of the skeleton words, we generate the skeleton words $\hat{S}$ to condition the second stage of our model. $C$ can be used to decode both the skeleton words and the area of the bounding boxes to fuse positional information with visual features, (Lu et al., 2019a), and 2) encode each feature vector with a Linear-ReLU-LayerNorm-Lineair instead of Linear embedding layer, where LayerNorm is layer normalization (Ba et al., 2016).

**Dual Staged Modeling:** In this approach, we introduce an intermediate natural-language interpretable skeleton $\mathcal{S}$ between $\mathcal{I}$ and $\mathcal{C}$. This $\mathcal{S}$ is composed of a sequence of lemmas, using a subset of content words ($s_1, s_2, ... s_n$) from $c$, where $n < m$. This reduces the output complexity of $f_\theta : \mathcal{I} \rightarrow \mathcal{C}$ by simplifying and denoising the noisy $\mathcal{C}$ to $\mathcal{S}$. Hence, the task of IC is decomposed into the first stage of predicting skeleton concepts and the second stage of caption generation using the intermediate skeleton.

**Stage 1: Skeleton Prediction (Img2Ske):** The first stage ($f_\theta : \mathcal{I} \rightarrow \mathcal{S}$) is to predict one of the 4 variants of the skeleton words (from §3.1) from the images. We experiment with both classification and generation paradigm that respectively do not possess and possess linear conditioning of the predicted skeleton word on the following words. We observe that the generation based skeleton prediction results in skeleton words that co-occur in a sentence. In contrast, the classification approach predicts skeleton words relevant to an image like person, man, singer that do not necessarily co-occur in a caption. This is detailed in §D of Appendix.

To improve co-occurrence of the predicted skeleton words, we generate the skeleton words $\mathcal{S}$ autoregressively where each word is conditioned on the previously predicted skeleton word. This conditional dependence models word co-occurrence more tightly as $p(\hat{s}_j | I, \hat{s}_{<j})$, making the skeleton a sequence of words. The model is optimized with cross-entropy loss, trained using teacher forcing. An attractive property is that the same architecture can be used to decode both the skeleton $\mathcal{S}$ and the caption $\mathcal{C}$. Moreover, the output tokens predicted in this stage are interpretable, and they are used to condition the second stage of our model.

**Stage 2: Skeleton-based Caption Generation:** The second stage of training uses images and skeletons to generate captions $f_\phi : \mathcal{I}, \mathcal{S} \rightarrow \mathcal{C}$. We experiment with 3 variants of conditioning predicted skeletons via encoding, decoding and autoencoding as shown in the model architecture in Fig. 2. The inputs, outputs and decoder attention conditioning for each stage are compared in Table 1.

| Input | Output | Conditioning |
|-------|--------|--------------|
| SkeEnc | $\mathcal{I} \rightarrow \mathcal{S}'$ | $h \rightarrow \mathcal{S}$, $C$ |
| SkeAE | $\mathcal{I} \rightarrow \mathcal{S}'$, $\mathcal{C}^+$ | $l \rightarrow \mathcal{S}'$, $\mathcal{C}^+$, $C'$ |
| SkeDec (no Stage 1) | $\mathcal{I} \rightarrow \mathcal{S}'$, $\mathcal{C}^+$ | $l \rightarrow \mathcal{S}'$, $\mathcal{C}^+$, $C'$ |

Table 1: The inputs and outputs of the different models.

In iterative refinement, $\mathcal{S}'$ is replaced by $\mathcal{C}'$.

2a. SkeEncoding: The predicted skeleton from the previous stage is used as input to the encoder. The image encoding and skeleton embeddings are fused with a unidirectional attention mechanism, called text-as-side (notated as $g$). In other words, we use the text representation as “side information” — each transformed image feature unit can attend to other image feature units (self-attention) and text (cross-attention), but text cannot attend to image. As shown in Fig. 2, this model has the dotted box in the Transformer encoder side, with the textual query, key, value ($Q_w$, $K_w$, $V_w$) and the visual counterpart attending to textual or visual key and value ($K_v+K_w$, $V_v+V_w$) with a visual query ($Q_v$). We focus on the text-as-side attention mechanism as our preliminary results indicate that it leads to qualitatively better captions than image-text co-attention (Lu et al., 2019b).

2b. SkeDecoding: The skeleton and caption are concatenated and predicted by the same decoder. This is not a two-staged model, as the model is trained to predict both skeleton and caption autoregressively. The model first predicts the skeleton words conditioned on the previously generated skeleton words, and then every token in the decoded caption attends to the entire predicted skeleton as well as the tokens of the caption decoded until that time step. The dotted box in Transformer decoder of Fig. 2 depicts this approach.

2c. SkeAE: To bring both the above models together, we simultaneously encode and decode the predicted skeleton. This brings the benefits of bidirectional attention on the input features (image and predicted skeleton words) and autoregressive attention on the re-predicted skeleton words while generating the caption. In this case, both the dotted boxes on encoder and decoder sides in Fig. 2 are active. The encoding mechanism follows the
function and the decoder prepends the caption generation task with the predicted skeleton.

4 Experiments and Results

Hyperparameters: Our transformer model uses 6 encoder and 6 decoder layers (unless specified otherwise), with 8 heads for multiheaded attention. Captions are subword-tokenized with a vocab size of 8,300. The models are optimized with Adam and an initial learning rate of $3 \times 10^{-5}$. We use mini-batches of size 128, and train for 1M steps. The token embedding and filter sizes are both 512. We experimented with several values for both frequency thresholding for skeleton words at 20, 50, 100 and k at 2, 4, 8, 16 for top-k selection for multilabel classification model. We manually selected the values that optimize the model performance based on manual examination as conducting human evaluations with more hyperparameters is very expensive especially with unreliable automatic metrics.

4.1 Datasets

Conceptual Captions (CC): CC (Sharma et al., 2018) is a large-scale dataset of 3.3M image-caption pairs covering a large variety of processed alt-texts from the web. The focus of this work is on denoising noisy captioning datasets (web-scale, not human verified). Hence our experiments are focused on CC, which is a step closer to having large and diverse alt-texts from the web at the cost of being noisy. In contrast, other popular datasets like COCO (size 120K) (Lin et al., 2014) and Multi30k (Elliott et al., 2016) are hand-annotated by humans and contain high quality images/captions. As a resource, CC is useful both for measuring progress on large-scale automatic captioning (Sharma et al., 2018; Changpinyo et al., 2019; Alikhani et al., 2020; Thapliyal and Soricut, 2020), as well as pre-training data for a variety of vision-and-language tasks (Lu et al., 2019b; Chen et al., 2020c; Tan and Bansal, 2019; Su et al., 2020; Li et al., 2020).

Pre-processing: CC might contain a long tail of spelling errors and other typos due to the automatic curation of the data. Therefore, we perform frequency based thresholding of the skeleton words to abate this noise. We experimented with several values for this hyperparameter and selected a minimum occurrence count as 50 that provides the desired balance between noise and vocabulary size.

|           | Iterative Refinement | Classification | Generation |
|-----------|----------------------|----------------|------------|
| Precision | 35.75                | 23.22          | 36.66      |
| Recall    | 24.29                | 41.31          | 24.30      |
| F-score   | 28.92                | 29.73          | 29.23      |

Table 2: Performance of skeleton prediction stage. Note that for classification and generation, the skeleton type used is ‘nouns & verbs’.

Multilingual CC: To demonstrate the cross-lingual transferability of our skeletons, we use automatic caption translations\(^2\) for CC, similar to the approach in (Thapliyal and Soricut, 2020). Note that the skeletons are learned from, and predicted in, English (not in the final target language), making English skeleton act as an interlingua. Since multilingual captions are all pivoted on English skeletons, this nullifies the requirement to 1) collect large-scale image-caption pairs in various language, and 2) have access to linguistic tools to analyze captions in each language. We perform experiments on 5 languages – French, Italian, German, Spanish and Hindi – which vary in word orders and token overlap with the English skeletons.

Conceptual Captions T2 test set: For human evaluations across all languages, we use T2 test set used in the Conceptual Captions Challenge\(^3\). It comprises of 1K out of domain images from the Open Images Dataset (Kuznetsova et al., 2020).

4.2 Automatic Evaluation

Skeleton Prediction: The goal of this stage is to extract key skeleton words from the image. We compute precision, recall and F-score as shown in Table 2. With the same labels (skeleton: nouns & verbs), both classification and generation approaches have similar F-scores. However, precision is higher for generation and recall is higher for classification based predictions. Based on both qualitative observations and human judgements, we note that generation approach was better, which shows that a higher precision is favorable in comparison to recall for this stage. The label size (of skeletons) in Table 2 is approximately 5K.

Skeleton-based Caption Generation: We report multilingual IC performance of baseline and our dual-stage models using CIDEr in Table 3 (English) and Table 4 (multilingual). Automatic metrics for captioning are based on surface n-grams, and are not suitable to evaluate when the ground truth cap-

\(^2\)We use the Google Cloud Translate API.

\(^3\)http://www.conceptualcaptions.com/
Table 3: Automatic metrics to compare various skeleton forms. Img2Cap is the baseline (large version refers to 12 encoder and decoder layers). Note that these results use generation-based skeleton prediction.

Table 4: CIDEr scores for skeleton (form: Nouns & Verbs, prediction approach: generation) conditioned caption generation for multiple languages.

Table 5: Ablations on val data for unpaired captioning.

Table 6: Human evaluation scores of different approaches and skeletons on English (vs the Img2Cap baseline).

4.3 Human Evaluations

Automatic metrics often have been found not to correlate well with human scores (Kilickaya et al., 2017; Alikhani et al., 2020) and do not correlate with human evaluations (§4.3). All the 4 proposed skeleton variants are evaluated systematically for automatic metrics, as shown in the last column of Table 3. However, since the automatic scores are compared against a gold standard of noisy captions, they are not reliable. Hence we conducted manual evaluation to select the best performing skeleton variant. Out of the 4 skeleton variants, nouns and verbs performed better in denoising and hence we demonstrated results for this variant for the remainder of the paper. We conducted further experimentation on nouns and verbs on the three models of dual staged captioning, controllability and cross-lingual transferability.

Multilingual captioning: Note that the skeletons are always in English, trained using annotations over the original English CC dataset. Cross-lingual results on val data of Multilingual CC are presented in Table 4. In addition to the data noisiness, a reason for slightly lower performance for non-English captions is probably noisy translation artifacts. For example, corresponding caption in the Hindi dataset for English caption “She is gazing at the fall colors” is वह घिरते रंगो की ओर देख रही है” (translation: She is looking at the falling colors.) Translation errors (such as ‘fall’ colors to ‘falling’ colors) introduce noise in the non-English datasets. Figure 3 presents an example of output multilingual captions for the baseline and our SkeAE approach.

Unpaired Image Captioning: A natural extension to our approach is for the caption generator to rely purely on predicted skeleton, and not use image features. This is a harder problem, but eliminates altogether, the need for image-caption pairs because the second stage (skeleton to caption) can be trained on a large text-only corpus. In this direction, within the scope of CC dataset, we investigate 1) with and without using image features in the second stage, 2) using ground truth skeleton (GTSke) to get an estimate of the upper bound on unpaired captioning 3) comparing the upper bound to the predicted skeleton (PredSke). These results are presented in Table 5. When image features are ignored, CIDEr drops by only 8 points when only predicted skeletons are used for caption generation compared to the baseline. This initial result shows that skeletons are a promising direction towards unpaired captioning.
Figure 3: Captions generated by baseline and our dual staged approach in 6 languages and their corresponding translations.

Figure 4: Human evaluation interface: We ask raters to: 1) compare the two captions (relative), 2) give ratings for each caption (absolute). Human annotators are asked to indicate the better caption relevant to the image.

| Language | Wins | Losses | Gains |
|----------|------|--------|-------|
| French   | 31.43| 29.53  | +1.9  |
| Italian  | 26.13| 24.93  | +1.2  |
| German   | 35.23| 33.93  | +1.3  |
| Spanish  | 34.03| 34.33  | -0.3  |
| Hindi    | 33.13| 28.63  | +4.5  |

Table 7: Human evaluation results for skeleton (form: nouns & verbs, prediction approach: generation with SkeEnc) conditioned caption generation for multiple languages.

**Results:** Table 6 presents the human ratings for English captions using different skeletons. From this, we observe the following:

(a) **Dual Staging helps:** Our dual staged models with skeletons (SkeEnc, SkeDec, SkeAE) show gains compared to the improved baseline Img2Cap model. Most notably, it shows that the ‘Nouns & Verbs’ skeletons significantly improves SkeEncoding model attaining the most significant gain, followed by SkeAE and then SkeDecoding.

(b) **Subselecting content words helps:** Using the same dual staged SkeEnc model without subselecting content words in the form of iterative refinement does not show any improvement in performance, supporting the hypothesis that sub-selecting content skeleton from noisy captions improves the overall caption quality.

(c) **Cross-lingual skeleton transfer:** Table 7 presents our human evaluation scores for captions in other target languages. We observe gains from the skeleton-based approach for 4 out of 5 languages, and only a slight loss for the fifth language i.e., Spanish, showing the effectiveness of cross-lingual transferability of the skeleton words. Our speculation for this is probably due to the dialect differences. The translation model that we used for Spanish is a mix of Spain Spanish and Latin American Spanish, with Latin American Spanish dominating. The evaluation was done by raters from Spain. The dialects are sufficiently different that it would impact the absolute scores.
To understand where the improvements quantified from the image. As observed in §4.3, as SkeAE skeleton predictor is trained to predict nouns and verbs captions of Visible skeleton-based IC results in the generation of more the image, such as 'fifth anniversary'. implications that the caption describes some potentially how/when/where the image was captured, such as responds to a caption containing details regarding a Visible age, such as 'people' or 'cake', is annotated using describing visually recognizable aspects of the im-
tween image-caption pairs. For instance, a caption

### 4.4 Cross-modal Discourse Coherence

To understand where the improvements quantified in Table 6 come from, we turn to the notion of discor-
coherence. Alikhani et al. (2020) introduce multimodal discourse coherence relationships be-
tween image-caption pairs. For instance, a caption describing visually recognizable aspects of the image, such as ‘people’ or ‘cake’, is annotated using a Visible relation; in contrast, a Meta relation cor-
responds to a caption containing details regarding how/when/where the image was captured, such as in ‘warm summer afternoon’, while a Story relation implies that the caption describes some potentially non-visible context behind the scene depicted in the image, such as ‘fifth anniversary’.

We hypothesize that our multi-stage approach of skeleton-based IC results in the generation of more captions of Visible type, as the intermediate skeleton predictor is trained to predict nouns and verbs from the image. As observed in §4.3, as SkeAE model performs better compared to the SkeEncoding and SkeDecoding models, we analyze the down-
stream captions based on SkeAE architecture. To assess this effect, we train the relation classifier de-
scribed in Sec. 4 of (Alikhani et al., 2020), and ob-
tain discourse relation labels for captions generated on T2-test images, by both the baseline Img2Cap and our SkeAE models. Table 9 (Counts columns) quantifies the shift of relation label distribution towards the Visible coherence relation, confirming our hypothesis. We also study the breakdown by coherence relations using the results from our hu-
man evaluations on the English captions. Table 9 (Human Evals column) reports this breakdown, in-
dicating that, of the 11.01% gains on human evals from Table 6, the shift from non-Visible to Visi-
ble discourse captions is associated with clear in-
creases in preference from the human raters. This is attributable to the fact that human raters are more likely to prefer captions that are in a Visible relation with the image, and therefore the shift towards generating Visible-type captions can be positively quantified in terms of human preference.
Controllability: Qualitative Discussion

The dual-stage modeling using skeleton decomposition can be a double-edged sword: it can be an information bottleneck, limiting the ability to train the model in an end-to-end manner; but, it brings forth the advantage of increased interpretability and thereby the ability to use the intermediate stage results to control the final caption. We present aspects of caption controllability by altering the skeleton to explore effects on caption length, informativeness, and gender specificity. This section discusses the utility of this dual staged model for controllability qualitatively with SkeAE architecture. Automatic intervention at the skeleton level involves non-trivially selecting related concepts for each image, and we leave this for follow-up work. Instead, we present an empirical study only to semi-automatically control gender specificity in two of the languages. We plan to conduct experiments to compare with other models (Zheng et al., 2019; Chen et al., 2020b) focused on controllability for follow-up work.

Effect of length of skeletons on captions: For applications that limit the caption lengths due to UI restrictions, the ability to control the length is important. The length of the skeleton correlates with the number of caption words, as shown in Figure 6. For 2 or 3 skeleton words, the percentage of captions monotonically decreases with the number of caption words, with the mode at 4-word captions. Thus, for skeletons of size 2, captions of length 4 are much more frequent than captions of length 6 or 8. For longer skeletons, we see that the mode shifts to the right: with skeletons of size 5, the caption length peaks between 8 and 10 words. Fig 7 illustrates this qualitatively.

Effect on gender specificity: Current models often make embarrassing mistakes when generating captions that mention gender. The availability of a skeleton provides a direct handle for human-in-the-loop correction of such biases, at a pre-caption-generation stage. This is more robust compared to caption post-processing, especially for highly inflected languages. To illustrate this, we compare the number of times ‘man’ appears in the captions generated by our baseline versus our dual-stage model after automatically modifying the skeleton (replacing ‘man’ to the gender-neutral word ‘person’ in the skeleton). Over the T2 dataset, the baseline caption generates ‘man’ 13 times, and the automatic control mechanism via our model reduces this by 46% (to 7 occurrences) in English. In Hindi, the equivalent of ‘man’ (आदमी) is generated 10 times, and it is reduced to a gender neutral word (व्यक्ति) by 70% (to 3 occurrences).

Effect of guiding information through skeleton: The skeleton acts as a knob enabling the model to describe different attributes of the image in the caption. Figure 5 presents an example of how varying the skeletons for two different images affect their captions. The words highlighted in green are derived from the skeleton and the ones highlighted in blue are image-related words.

6 Conclusions

Scaling image captioning models practically mandates training on noisy and uncurated data available on web. Our work presents an approach that denoises learning from such large yet diverse web-scaled data with alt-text annotations by sub-selecting content as intermediate skeletons. We experimentally demonstrate that this approach improves the captions significantly in human evaluations on out-of-domain test data by converting meta and story like captions to more visually informative captions. We also demonstrate the transferability of English skeleton words to improving captions in five other languages. Additionally, the natural-language interpretable skeleton layer gives us a way to better control and perform human-in-the-loop corrections of model predictions. We believe that this is a promising direction towards unpaired IC and also has potential for semi-automatic interventions to correct or interact with the skeletons to guide the final captions.

In this work, our main focus is denoising alt-text captions using skeletons and using them for cross-lingual captioning. In future, we plan to explore the effect of denoising in pretraining large multimodal models (BLIP (Li et al., 2022), UNITER (Chen et al., 2020d), ViLBERT (Lu et al., 2019c)) as base architectures by automatically cleaning captions, similar to how BLIP has an additional classifier to subselect captions that are not noisy. Appendix H presents a broader impact of our work.
Jiuxiang Gu, Shafiq R. Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 10322–10331. IEEE.

Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Haqing Lu. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4204–4213.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In Proceedings of NIPS workshop.

Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shihatuddin, and Hamid Laga. 2018. A comprehensive survey of deep learning for image captioning. CoRR, abs/1810.04020.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In CVPR.

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.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2020. Ultra fine-grained image semantic embedding. In Proceedings of the 13th International Conference on Web Search and Data Mining, pages 277–285.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers, pages 199–209. Association for Computational Linguistics.

Suwon Kim, HongYong Choi, JoongWon Hwang, JangYoung Song, SangRok Lee, TaeKang Woo, and AI Modulabs. 2019. Vizwiz image captioning based on aoanet with scene graph. ivc.ischool.utexas.edu.

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

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. volume 35, pages 2891–2903. IEEE.

Alina Kuznetsova, Mohamad Hassan Mohamad Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale.

Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. 2014. TreeTalk: Composition and compression of trees for image descriptions. Transactions of the Association for Computational Linguistics, 2:351–362.

Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In Proceedings of the 25th ACM international conference on Multimedia, pages 1549–1557.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. CoRR, abs/2004.05569.

Gen Li, Nan Duan, Yuejian Fung, Daxin Jiang, and Ming Zhou. 2020. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In AAAI.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019a. Entangled transformer for image captioning. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 8927–8936. IEEE.

Jiangyun Li, Peng Yao, Longteng Guo, and Weicun Zhang. 2019b. Boosted transformer for image captioning. Applied Sciences (2076-3417), 9(16).

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. CoRR, abs/2201.12086.

Nannan Li and Zhenzhong Chen. 2018. Image captioning with visual-semantic LSTM. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, pages 793–799. ijcai.org.

Siming Li, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 220–228.
Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer.

Zachary C. Lipton. 2018. The mythos of model interpretability. Commun. ACM, 61(10):36–43.

Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. 2018. simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 137–149.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In NeurIPS.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In NeurIPS.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019c. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 13–23.

Ruotian Luo and Greg Shakhnarovich. 2020. Controlling length in image captioning. CoRR, abs/2005.14386.

Alexander Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8591–8600.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 91–99.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. IJCV, 115(3):211–252.

Paul Hongsuck Seo, Piyush Sharma, Tomer Levinboim, Bohyung Han, and Rudu Soricet. 2020. Reinforcing an image caption generator using off-line human feedback. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020. The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020. The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 2693–2700. AAAI Press.

Himanshu Sharma, Mannmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. 2020. Image captioning: A comprehensive survey. In 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), pages 325–328. IEEE.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricet. 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.

Wejie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jiefeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In ICLR.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers in EMNLP-IJCNLP.

Ying Hua Tan and Chee Seng Chan. 2016. phi-lstm: A phrase-based hierarchical LSTM model for image captioning. In Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V, volume 10115 of Lecture Notes in Computer Science, pages 101–117. Springer.

Ashish V. Thapliyal and Radu Soricet. 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, ACL 2020, Online, July 5-10, 2020, pages 160–170. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 963–969. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In Proceedings of
Satoshi Tsutsui and David Crandall. 2017. Using artificial tokens to control languages for multilingual image caption generation. arXiv preprint arXiv:1706.06275.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Dalin Wang, Daniel Beck, and Trevor Cohn. 2019a. On the role of scene graphs in image captioning. In Proceedings of the Beyond Vision and LANguage: inTEGRating Real-world kNowledge (LANTERN), pages 29–34.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019b. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE International Conference on Computer Vision, pages 5310–5319.

Yue Zheng, Yali Li, and Shengjin Wang. 2019. Intention oriented image captions with guiding objects. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 8395–8404. Computer Vision Foundation / IEEE.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale japanese image caption dataset. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers, pages 417–421. Association for Computational Linguistics.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4651–4659.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. arXiv, 1905.07841.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 11–20. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057.

Xu Yang, Kaifeng Tan, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10685–10694.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In Proceedings of the European conference on computer vision (ECCV), pages 684–699.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In Proceedings of the IEEE International Conference on Computer Vision, pages 4894–4902.
A Comparison of SkeEnc and SkeAE on multilingual captions

We have discussed the human evaluation scores of the SkeAE model by using nouns and verbs as skeletons in Table 7 in the main paper. In addition to this, we also conducted human evaluation to compare the SkeEnc model with the nouns and verbs skeletons in comparison to the baseline. We present this in Table 10. While there are improvements in 3 languages, the performance is also hurt in two languages. However, as we see, by comparing the performances in Table 7 and Table 10, we observe that SkeAE has a clear advantage when leveraging the English caption to improve multilingual captions. This clearly indicates that channelling the prediction of the skeleton words in conjunction with the caption itself is enabling the model decoder to attend to the previously predicted skeleton words in the same decoder.

| Language | Wins | Losses | Gains |
|----------|------|--------|-------|
| French   | 31.93| 31.43  | +0.50 |
| Italian  | 33.13| 28.32  | +4.81 |
| German   | 29.43| 29.72  | -0.30 |
| Spanish  | 30.53| 34.43  | -3.90 |
| Hindi    | 29.93| 26.03  | +3.90 |

Table 10: Human evaluation results on SkeEnc model for skeleton (form: nouns & verbs, prediction approach: generation) conditioned caption generation for multiple languages.

B Comparison of Classification and Generation based Skeleton Prediction

From a preliminary manual analysis, we observed that the classification based approach to skeleton prediction faces the problem of predicting words that are related but are not likely to co-occur within the same sentence in the caption. This is described in detail in points 1a and 1b of §3. To validate this observation, we conducted human evaluation of the captions generated from classification and generation based approaches relative to one another. This setup is different from the rest of the experiments in human evaluation in the paper which compare any given model relative to the baseline model. In contrast, this study is to compare the generation and classification approaches with one another. These results are presented in Table 11.

The top-8 highest scoring content words are chosen to reduce input noise for the caption generator while improving the recall of concepts. We experimented with different values for this and selected 8 to be an optimal balance between the content in the skeleton words and the noise.

We observe that the generation based approach has significant gains of +8.91 over the classification based approach. Most of the prior literature uses the classification based approach to predict content or bag of concepts to assist caption generation. Our hypothesis is that this classification based model helps in end-to-end approaches where the loss from caption generation backpropagates to the classifier model as well. As opposed to this, our model decouples the prediction of the skeleton or concept words that are further used for caption generation. Hence we believe that suppressing the words that do not co-occur is important in the skeleton prediction task and the generation based approach is addressing this problem.

C Absolute Ratings

Here are some of the observations from these results:

• Better results of Dual Staged Approach: As we can see in the last two rows (rows 8 and 9), our proposed SkeEnc and SkeAE show absolute improvements in both the categories. This further demonstrates that the proposed dual staged approach is generating better denoised captions when trained on noisy uncurated alt-text–based captions.

• Sub-selecting content words is better: Now that we saw the improvements with the dual staged approach, we now investigate whether sub-selecting content words is important. For this, we present comparison between rows 7 and 8. Both these models are dual staged with SkeEnc i.e encoding the predicted skeleton in the second stage. The only difference is that row 8 sub-selects all nouns and verbs to predict the skeletons whereas row 8 includes all the words from the captions to predict the skeletons. Row 8 shows better performance compared to row 7. This means that sub-selecting content words contribute to the caption generation in the second stage.
Please note that we focus on alt-text based captions, so we experiment on Conceptual Captions instead of cleaner alternatives such as MSCOCO and Multi30k. The latter do not include as noisy captions as they are hand-annotated (refer Section 4.1)

D Img2Ske: Classification based prediction

Skeleton prediction is posed as a multilabel classification problem where the prediction of a skeleton word $s_i$ is not conditionally dependent on the prediction of another skeleton word $s_j$. Our goal is to evaluate the effectiveness of simple generation and classification models to predict skeletons, and naturally generation based approach reduces redundancies due to conditional dependence of label/skeleton prediction. The encoder part remains the same as the baseline followed by optimization with sigmoid cross entropy between the skeleton words $S$ and image encoding $z_I$, which is the representation of the image from the encoder.

Accuracy, $A = \frac{1}{N} \sum_{i=1}^{N} \frac{|S_i \cap \hat{S}_i|}{|S_i \cup \hat{S}_i|}$ (1)

The skeleton for the second stage is chosen as the ordered list of top-8 (experimentally selected) high scoring words after the softmax layer. However, conditional independence of skeleton words with one another ignores the co-occurrences of words capable of composing a sentence or a final caption. For instance, classification predictions are composed of words and their synonyms that are highly correlated like \{person, man, singer\}. These words definitely are relevant to an image but do not all necessarily co-occur in a sentence.

Table 2 presents the precision, recall and f-scores of the generation and classification based approaches for skeleton prediction. These metrics, however are misleading because they do not account for synonyms or semantic similarity. For example, ‘food’, ‘meal’, ‘lunch’ and ‘dinner’ are all distinct labels while computing these metrics, and predicting one instead of the other get heavily penalized even though the effect on downstream caption quality would be minimal. This issue gets amplified by the fact that with CC that has a rich vocabulary with words such as electricity ‘pylon’ and ‘tower’ referring to the same concept.

E Performance drop for Spanish

While we have seen improvements in the performance on multiple languages in human evaluation (Table 6), we observed a drop in the preference for Spanish captions when we use skeletons. Given the similarity in word order between Spanish and English in comparison to Hindi, the lower performance of Spanish is an interesting result indeed. Our speculation for this is probably due to the dialect differences. The translation model that we used for Spanish is a mix of ‘Spain Spanish’ and ‘Latin American Spanish’, with Latin American Spanish dominating. The evaluation was done by raters from Spain. The dialects are sufficiently different that it would impact the absolute scores.

F Intuition for skeleton words:

The alt-text captions are silver standard and harbor a lot of diversity. Hence filtering frequently occurring words based on a frequency cutoff as the skeletons helps balance between conditioning on the frequent words (not noise) and diverse concepts. Qualitatively, consider an image of a house with the caption ‘apartment for rent’ and ‘apartment for sale’. With the frequency based skeleton selection, the noun word ‘apartment’ is selected as skeleton ignoring the rest. In this way, we are denoising alt-text captions to generate captions with visible concepts.

G Hyperparameters:

This section lists the hyperparameters used for training our models. We used BERT embeddings (Devlin et al., 2019) to initialize the words in skeletons in the SkeEnc and SkeAE models.

- Learning rate: We experimented with $3.2e^{-5}$, 0.5, 1, 1.5 and 2 as the learning rate. The experiments presented in the paper have the learning rate of 1. The learning rate is decayed at 0.95 decay rate with staircase strategy.
- Number of layers: All our models have 6 layers for encoder and decoder. We also conducted an additional experiment to check if the model complexity of the end-to-end baseline can improve the performance in comparison to our dual staged approach. To evaluate this, we doubled the number of layers where the number of transformer encoder and decoder layers are 12 each as presented in the
paper as Impr Img2Cap (large) in Table 3 in Section 4.2.

- **Subtoken Vocabulary:** We experimented with 4000 and 8300 sub-token vocabularies. The experiments in the paper all have 8,300 as subtoken vocabulary size.

- **Batch size:** All our experiments include batchsize of 128 only.

- **Number of steps:** We train for a maximum of 1 million update steps.

- **Maximum Caption Length:** In the baseline and the SkeEnc models, our decoder generates a maximum words of length 36. In the SkeAE and SkeDec model, the skeleton words are prepended to the caption. So we allow the decoder to generate 72 words in these two models.

- **Warm up and decay steps:** The model is warmed up for 20 epochs and decayed for 25 epochs.

- **Embedding size:** We use embedding dimension of 512.

- **Beam size:** We perform beam search in the decoder with a beam size of 5.

Here are some of the configuration and modeling choices for training the models:

- **Attention type:** Our experiments include attention types of cross-attention and text-as-side as described along with point 2a in Section 3.

- **FRCNN Tokens:** We use 1601 tokens from the trained FRCNN.

## H Broader Impact

We believe that this work has extensive impact in scaling captioning models to large and noisy datasets thereby exploiting web data and reduce manual annotation efforts. We do not foresee any immediate concerns ethically directly from our work. However, while applying this to datasets crawled from the web, offensive content should be removed. In general, we envisage researchers and practitioners to benefit from our approach especially, when expensive human annotations are not available. More broadly speaking, we also strongly believe that our approach laid blocks for future work on cross-lingually leveraging English skeletons and automatic translations to generate captions for various languages. Hence, when combined with unpaired captioning, this can especially benefit captioning in low resource languages.