ICC: Quantifying Image Caption Concreteness for Multimodal Dataset Curation

Anonymous ACL submission

Abstract

Web-scale training on paired text-image data is becoming increasingly central in multimodal learning, but is challenged by the highly noisy nature of datasets in the wild. Standard data filtering approaches succeed in removing mismatched text-image pairs, but permit semantically related but highly abstract text. In this work, we propose a new metric, *Image Caption Concreteness (ICC)*, that evaluates caption text without an image reference to measure its concreteness and relevancy for use in multimodal learning. Our approach leverages strong foundation models for measuring visual-semantic information loss in multimodal representations. We demonstrate that this strongly correlates with human evaluation of concreteness in both single-word and sentence-level texts. Moreover, we show that curation using *ICC* complements existing approaches and succeeds in distilling multimodal web-scale datasets for more effective learning.

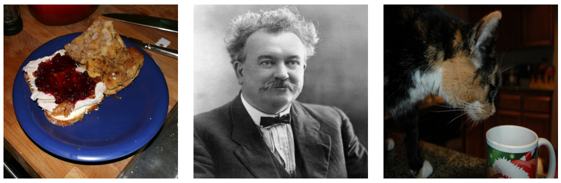


Figure 1: Given an image caption, *ICC* measures its visual concreteness. We show samples from MS-COCO (Lin et al., 2014), containing captions with low (↓) and high (↑) *ICC* scores. As illustrated, our method detects highly abstract captions, which are problematic in the context of multimodal learning. It does so by learning to quantify visual-semantic information loss in multimodal foundation models.

1 Introduction

Pre-training large vision-language models (VLMs) on web-crawled datasets consisting of image-caption pairs has become the standard practice in achieving state-of-the-art results in vision-and-language tasks such as image captioning and multimodal representation learning. However, raw web data are often noisy and contain many low-quality samples, which impair VLMs' learning in terms of quality and efficiency (Li et al., 2022; Schuhmann et al., 2022; Radenovic et al., 2023). While various factors impact data quality, we focus on *semantic* noise, characterized by analyzing the meaning of data items rather than, e.g., identifying low resolution images or quantifying token repetitions.

Existing datasets are commonly filtered using VLMs such as CLIP (Radford et al., 2021) to identify image-text semantic misalignments (Sharma et al., 2018; Schuhmann et al., 2022), namely, captions irrelevant to their images, or using rule-based proxies such as measuring the complexity of captions via semantic parsing (Radenovic et al., 2023). However, these approaches fail to identify captions that are highly abstract and may contain subjective, non-visual information, despite being semantically aligned with the image and having a sufficiently complex grammar. Figure 1 shows examples of such image-caption pairs. A caption such as *"It does not look like something I would want to eat"* is semantically related to the image, but a model trained to predict this caption from its image may learn to hallucinate details, e.g., liking a certain type of food in this example, which are not visually grounded and are highly subjective.

In this vein, we consider the *visual concreteness* of image captions, referring to the degree to which text describes a specific visual scene that can be vividly imagined (as opposed to abstract text that may correspond to many possible visual representations). Visual concreteness provides a comple-

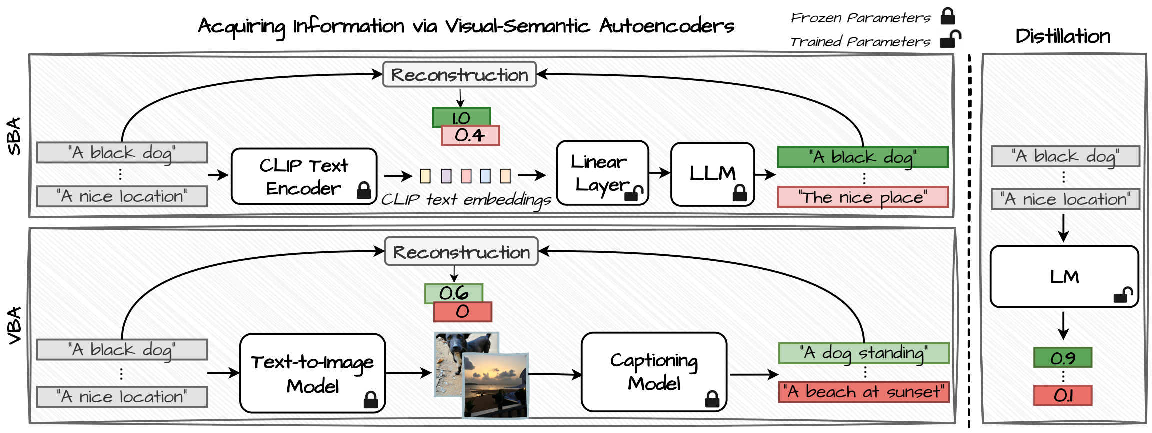


Figure 2: Predicting visual concreteness scores of image captions with our method. We first acquire information using a semantic-bottleneck autoencoder (SBA, top left) and an visual-bottleneck autoencoder (VBA, bottom left). We then distill a weighted combination of their reconstruction scores into a smaller language model (LM, right), which learns to produce *ICC* scores for new texts. We visualize reconstruction scores for highly concrete ("A black dog") and highly abstract ("A nice location") texts. High and low scores are colored in green and red, respectively. As illustrated, our final score, which combines the two pipelines, yields more accurate concreteness predictions.

mentary dimension of textual quality to consider for vision-and-language tasks, as filtering captions by concreteness is a natural way to encourage visually-grounded predictions.

We propose the *Image Caption Concreteness (ICC)* metric for quantifying the visual concreteness of image captions calculated from text alone, i.e., without an image reference. We measure concreteness using autoencoding pipelines with visual-semantic information bottlenecks, previously used for other aims (Kamath et al., 2023; Yang et al., 2023). Specifically, we use a semantic-bottleneck autoencoder that identifies how well an LLM recovers the input caption from its semantic CLIP embedding, and a visual-bottleneck autoencoder that leverages the competence of text-to-image generative models. Our *ICC* metric is distilled from these pipelines; see Figure 2.

Extensive experiments show *ICC*'s effectiveness in filtering multimodal web-scale data for downstream tasks such as image captioning and text-based image retrieval. We will release our data, code, and trained models, anticipating the use of *ICC* for further tasks that require curation of web-scale visually-grounded text.

2 Method

Given an image caption (of an *unseen* image), we aim to predict its degree of visual concreteness. Our underlying assumption is that more visually concrete text can be mapped to or from a visual representation with less information loss. Conversely, we expect that visually abstract text cannot be converted to or from a visual representation without significant information loss, since it does not clearly describe a well-defined image. We model this process with autoencoder components that convert text to and from visual-semantic representations, and quantify the information loss of this process as a proxy for visual concreteness. We proceed to describe our proposed semantic-bottleneck autoencoder and visual-bottleneck autoencoder components, and their consolidated distillation into the *ICC* score.

Semantic-bottleneck Autoencoder (SBA). Motivated by findings that CLIP embeddings encode visual information in text and particularly concreteness (Alper et al., 2023), we construct an autoencoding pipeline with CLIP text embeddings as a semantic information bottleneck, as shown in Figure 2 (top left). We extract visual information from the CLIP text embedding space by utilizing a frozen LLM (`Llama-2-7b`, Touvron et al., 2023), training a linear layer that converts the VLM text encoder's output to inputs for the LLM. The training objective aims at reconstructing the input captions via a token-wise cross-entropy objective.

After training SBA over image–caption pairs, we use it for measuring text concreteness by encoding and decoding the text followed by measuring reconstruction fidelity via per-character Edit Distance (Levenshtein et al., 1966), normalized by caption length as detailed in the appendix. This pipeline succeeds in reconstructing highly concrete text (such as "A black dog" shown in Figure

	Word Conc.			Sentence Conc.		
Method	ρ	ρ_s	τ	ρ	ρ_s	τ
CLIP-SP	0.60	0.62	0.44	-0.36	-0.35	-0.27
aveCLIP	0.55	0.56	0.39	0.29	0.28	0.22
ICC	**0.75**	**0.75**	**0.55**	**0.69**	**0.67**	**0.54**

Table 1: **Concreteness evaluation on single-word and sentence-level texts**, measured using Pearson ρ, Spearman ρ_s, and Kendall (τ) correlation coefficients.

Data	B@4	M	R	C	S	BSc
CC	9.9	15.0	37.5	34.6	96	0.47
CC+CLIP	10.4	15.3	38.3	36.2	98	0.48
CC+CA	9.2	14.7	36.2	31.8	93	0.47
CC+*ICC*	**11.8**	**16.3**	**40.7**	**42.5**	**109**	**0.51**
LA	0.5	4.8	12.9	1.9	14	-0.04
LA+CLIP	0.2	4.0	11.0	1.1	9	-0.07
LA+CA	0.2	3.9	10.7	1.0	9	-0.07
LA+*ICC*	**7.8**	**12.2**	**30.5**	**21.2**	**72**	**0.35**

Table 2: **Captioning results using 500k filtered samples** over the MS-COCO Karpathy test split. *Data* denotes the training dataset – Conceptual Captions (CC) or LAION-400M (LA). We compare our performance (+*ICC*) to two filtering baselines: +CLIP indicates filtering by top CLIP similarity and +CA indicates Complexity and Action filtering. We also report performance obtained by randomly selecting 500k samples (1st and 4th rows). B@4, M, R, C, S and BSc denote BLEU-4, METEOR, Rouge-L, CIDEr, SPICE, and BERTScore metrics respectively.

	COCO			Flickr		
Data	R@1	R@5	R@10	R@1	R@5	R@10
LA	4.5	15.0	23.0	9.6	27.7	40.5
LA+CLIP	2.2	8.0	13.1	4.9	15.1	23.1
LA+CA	6.5	19.7	29.3	16.3	40.5	55.2
LA+*ICC*	**10.0**	**27.1**	**38.4**	**21.7**	**49.6**	**62.4**

Table 3: **Text-to-image retrieval results for representations** trained on 500k samples with different filtering methods: LA indicates 500k random samples from LAION-400M, +CLIP indicates filtering by CLIP similarity; +CA indicates Complexity and Action filtering.

2). However, while abstract captions are expected to yield generally poor reconstructions, their measurements are less consistent (e.g. "A nice location" yields a non-negligible reconstruction score of 0.4). To more robustly handle such cases, we propose our VBA component, detailed next.

Visual-bottleneck Autoencoder (VBA). The VBA is constructed by using images as an intermediate representation through which textual information passes. In particular, we concatenate a text-to-image model (Stable Diffusion 2, Ramesh et al., 2022) and a captioning model (`BLIP-2`, Li et al., 2023). For this pipeline, all components are frozen and no training is required; we directly measure information loss as a result of mapping to and then from images, as shown in Figure 2 (bottom left). Due to the difficulty of reconstructing exact matches from images, we measure the semantic fidelity in reconstruction (rather than edit distance) using BERTScore F1 score (Zhang et al., 2019).

***ICC* Score.** We assemble SBA and VBA reconstruction scores over a collection of image–caption pairs and distill their aggregated values into our final *ICC* score. Specifically, we train a small text encoder model (Liumm et al., 2019) over a linear combination of the two scores, with weights computed by regressing over a set of annotated captions. Additional details, ablations and visualizations are provided in the appendix.

3 Results and Discussion

We proceed to first show *ICC*'s correlation to concreteness (Section 3.1), followed by its benefit in data curation for downstream tasks (Section 3.2).

3.1 Concreteness Correlation

Table 1 shows the correlations of different concreteness estimation methods to ground-truth concreteness scores on both single-word and sentence-level (caption) benchmarks. We compare to zero-shot probing of CLIP through Stroop probing (SP) as proposed by Alper et al. (2023). We also compare to aveCLIP (Wu and Smith, 2023), which generates multiple images from a caption and measures the average similarity between the text and generated images. Due to its high computational cost, we only evaluate it on a statistically-significant portion of the single-word benchmark, which contains nearly 15K samples.

Correlation to Word Concreteness. We first validate our metric by measuring it on a dataset introduced by Hessel et al. (2018). This dataset is composed of 39,954 English uni-grams and bigrams coupled with human-labelled concreteness scores on a scale from 1 (abstract) to 5 (concrete), averaged over annotators. To compare with prior work, we only use unigram nouns, totaling 14,562 items. As illustrated in Table 1, *ICC* significantly outperforms prior works over all correlation metrics.

Correlation to Caption Concreteness. We manually annotated concreteness scores for 200 captions from LAION-400M (Schuhmann et al., 2022); see the appendix for more details. As Table 1 shows, our method exhibits superior correlation with human judgements of text-level concreteness, providing further motivation for its use.

3.2 VLM Dataset Curation

Captioning Models. In Table 2 we show quantitative results of applying *ICC* filtering on top of standard CLIP filtering over different datasets for training a captioning model. We hold the dataset size fixed for all experiments. The captioning model used is an encoder-decoder architecture with a pretrained Swin (Liu et al., 2021) vision encoder and GPT-2 (Radford et al., 2019) text decoder, trained for a single iteration on each training sample. Additional training details are provided in the appendix. We compare to two filtering methods – top-CLIP similarity filtering and Complexity and Action filtering (Radenovic et al., 2023), using our re-implementation, as there is no publicly-available code. The latter is a rule-based filtering method which aims to retain only sufficiently complex captions that also contain an action, based on semantic parsing. As illustrated in the table, filtering with *ICC* outperforms alternative filtering methods for captioning given a fixed number of desired samples and training iterations. As can also be observed in the table, filtering with a fixed CLIP similarity threshold may even *degrade* performance, suggesting that samples with very high CLIP similarity are not necessarily better for training captioning models.

Image-Text Representation Learning. We also perform a representation learning experiment by training a dual text and image encoder model on a dataset filtered with different methods. Table 3 reports performance over standard retrieval benchmarks, namely COCO (Lin et al., 2014) and Flickr (Plummer et al., 2015). The model is initialized from a pretrained vision-encoder (Dosovitskiy et al., 2010) and text-encoder (Devlin et al., 2018) as suggested by Zhai et al. (2022). All other experimental settings are identical to the captioning model training. As illustrated in the table, *ICC* yields superior performance for this task.

4 Related Work

Evaluating Text Concreteness. Word concreteness is a topic of interest in cognitive science (Schwanenflugel, 2013), and a number of works have studied automatic prediction of word concreteness using machine learning (Hill et al., 2014; Hill and Korhonen, 2014; Hessel et al., 2018; Rabinovich et al., 2018; Charbonnier and Wartena, 2019; Alper et al., 2023). However, little attention has been paid to measuring concreteness at the sentence or string level. Most similar to us is Wu and Smith (2023), who generate multiple images for each caption and average the similarities over all the images to produce a sentence-level concreteness score. Other text evaluation metrics compare to reference texts (Gehrmann et al., 2023) or a reference image (Hessel et al., 2021), while we are interested in the inherent quality of text in isolation (namely, its visual concreteness).

Multimodal Dataset Curation. Due to the highly noisy nature of Internet multimodal data, prior works have filtered using approaches such as rule-based text parsing (Radenovic et al., 2023), using CLIP similarity to detect misaligned text-image pairs (Schuhmann et al., 2022), and de-duplicating semantically similar content (Abbas et al., 2023). A number of prior works have also proposed replacing or augmenting multimodal datasets with synthetic samples (Li et al., 2022, 2023; Fan et al., 2023). By contrast, our approach does not require modification of the given dataset and identifies semantically infelicitous captions allowed by prior methods. Our work also contrasts with dataset distillation, which has been applied to multimodal dataset curation (Wu et al., 2023); while dataset distillation methods select samples to explicitly optimize a chosen downstream objective, we focus on the simpler and more general task of identifying samples of inherently poor quality.

5 Conclusion

We present a new metric for measuring the visual concreteness of image captions without an image reference. By leveraging strong foundation models, we quantify visual-semantic information loss and find that this highly correlates with human concreteness judgments. Our results demonstrate that *ICC* is effective at multimodal data filtering. We foresee the use of *ICC* in additional tasks requiring the curation of web-scale multimodal data, where visually concrete text is needed.

Limitations

While our method detects and filters an important category of noise in multimodal datasets, we note that abstract captions such as those in Figure 1 may contain important information which our method discards. Future work might instead extract the relevant visual information from such captions, to avoid losing the information signal in such items. We also note that such captions often contain external or subjective information which could be of interest to tasks such as news image captioning or multimodal sentiment analysis, where external context is of interest. To identify such cases, further work might enhance the interpretability of our method to explore *why* a caption is or is not concrete.

Ethics Statement

Models trained on multimodal Internet data may inherit biases from their training data. Our method is not designed to filter potentially harmful image descriptions; moreover, such biases are also present in the models used as part of our pipeline (CLIP, generative models) and thus our model may possibly inherit or amplify these issues for downstream tasks. We anticipate further research into such biases and guidelines needed before putting these models into deployment.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. 2023. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6778–6788.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 176–187. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2010. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *NAACL*.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 255–265.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders are performance bottlenecks in contrastive vision-language models. *arXiv preprint arXiv:2305.14897*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Yinhan Liumm, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

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.

Ella Rabinovich, Benjamin Sznajder, Artem Spector, Ilya Shnayderman, Ranit Aharonov, David Konopnicki, and Noam Slonim. 2018. Learning concept abstractness using weak supervision. *arXiv preprint arXiv:1809.01285*.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6967–6977.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Paula J Schwanenflugel. 2013. Why are abstract concepts hard to understand? In *The psychology of word meanings*, pages 235–262. Psychology Press.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Si Wu and David Smith. 2023. Composition and deformance: Measuring imageability with a text-to-image model. In *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 106–117, Toronto, Canada. Association for Computational Linguistics.

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. 2023. Vision-language dataset distillation.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023. Multicapclip: Auto-encoding prompts for zero-shot multilingual visual captioning. *arXiv preprint arXiv:2308.13218*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix A Implementations Details

A.1 Dataset Used For SBA & VBA

In all stages we use a subset of CC3M (Sharma et al., 2018), composed of 595k samples curated by Liu et al. (2023) to provide larger concept coverage. We use 80% of the samples for training the linear layer of the SBA, for a single epoch. We then split the remaining 20% again to 80%/20%. We then use the 80% for generating reconstructions of the SBA and VBA. These reconstruction scores are used as the labels for the distillation phase which is being done using the remaining 20%.

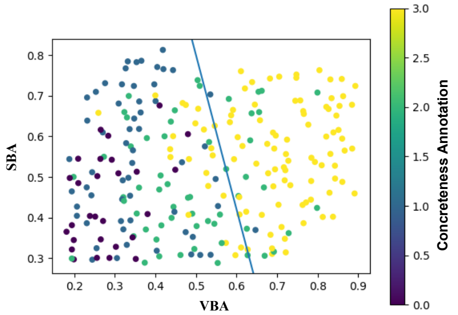


Figure 3: **Finding the Optimal Weights.** We measure the optimal combination of the two scores with respect to GT concreteness annotations.

A.2 Optimal Weighting of Scores

To find weights of the SBA and VBA scores for the final *ICC* distillation, we regress using logistic regression (where we label a caption as concrete if it is above the median score and abstract if it is below the median score) over a set of 244 samples captions, sampled uniformly over the VBA and SBA score, which we manually annotate with concreteness scores as shown in Figure 3. As seen in the figure, both scores contribute to the optimal predicted concreteness score. Note that the set of annotated captions used for selecting the SBA and VBA scores is separate from our manually annotated sentence concreteness benchmark used for calculating correlation scores, thus avoiding data leakage.

A.3 Normalizing By Caption Length

We aim to have reconstruction scores that are only dependent on the concreteness of captions and not on the length of the captions. In Figure 4, we show the distribution of the reconstruction similarities before and after normalization per caption length. We can see in Figure 4a that there is a strong dependency on caption length, which we would like to avoid.

More specifically, we force the reconstruction similarity distribution to be distributed according to $\mathcal{LN}(\mu = 0.5, \sigma = 1)$, where $\mathcal{LN}$ denotes a Logit-Normal distribution. The normalization is performed by standardizing the logit of the similarities (defined by $ln(\frac{1}{1-p})$) for each caption length, and then taking the inverse logit. We can see in Figure 4b that short captions are reconstructed more easily compared to longer ones, and that normalization by caption length successfully disentangles the reconstruction scores from the caption length dependency.

A.4 Datasets Used in Our Experiments

We use subsets of CC3M for training the captioning model and subsets from CC3M and LAION-400M for training the image-text representation model. For LAION, we only sample 8M samples, filtered with the provided NSFW filter to remove unsafe contents. For CC3M, we filter all samples with CLIP similarity below 0.3 (note that LAION-400M is already filtered with 0.3 threshold of CLIP similarity), leaving us with 1M samples. From these initial datasets, we further filter using the methods described in the main paper.

A.5 *ICC* Distillation

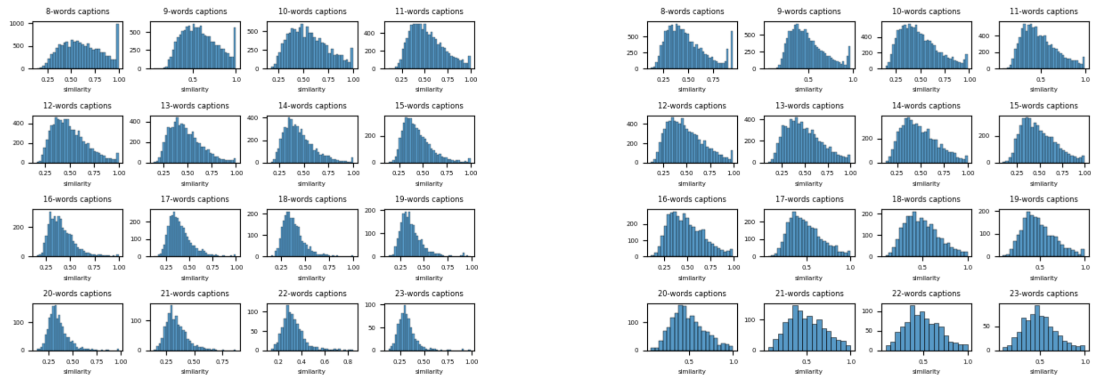
We distill the knowledge obtained by the two pipelines described in the paper in a two-stage manner. Firstly, we distill the VBA and SBA scores into two distinct DistilRoBERTa (Liumm et al., 2019) models. We then collect a small subset of 244 captions, sampled to have approximately uniform joint distribution of scores, and annotate the concreteness scores of these captions. This is showcased in Figure 3. We regress over these samples to get the optimal weights as discussed in A.2. We then use this optimal combination as the labels for training the final *ICC* model used for all the experiments in the paper. All distilled models are trained with a Mean Squared Error (MSE) objective.

A.6 Caption Concreteness Benchmark

Next we describe the data collection and annotation details. Our aim is to have a small, yet diverse set of samples that represent the wide diversity of possible captions. Since Laion-400M is very noisy and only a small portion of it includes highly concrete captions, we sample 150 items that satisfy the following rules:

- The caption must include at least 10 character
- The caption must not contain more the 80% of capitalized words.
- The caption must include at least 2 stop words, filtered using NLTK parser (Loper and Bird, 2002).
- The ratio of stop words to all the words in a captions must not exceed 20%.

The remaining 50 samples in our dataset are selected randomly to include more "raw" captions as well. For all captions in our benchmark, we also apply NSFW filtering and make sure the caption do not include offensive or personal content.



(a) Before Normalization

(b) After Normalization

Figure 4: **Normalizing by caption length.** We show the reconstruction similarity scores of SBA for each caption length before normalization (in 4a) and after normalization (in 4b).

We show all the captions in Figure 7, sorted according to the annotated concreteness scores. As illustrated in the figure, we were able to achieve a relatively good coverage of various abstraction levels using the aforementioned sampling process.

We note that LAION-400M has an open access license, and we will release our benchmark to facilitate further research in the direction of quantifying caption concreteness.

A.7 Zero-Shot CLIP Concreteness Score

We adapt the Stroop Probing method (Alper et al., 2023) that is originally designed to assess the concreteness of words, to captions. We follow the same procedure used when measuring concreteness of words, but replace the empty slot in the prompts with a caption rather than a single word, and use only the prompts that fit the context of caption in the black spot (i.e., we don't use the captions "Alice giving the [*] to Bob" and "Bob giving the [*] to Alice" as they aren't appropriate when using a caption to replace the empty slot [*]).

A.8 aveCLIP Word Concreteness

Since aveCLIP requires generating many images per word, we found that running aveCLIP over the entire word concreteness dataset is not feasible due to runtime constraints. Therefore, we sample 150 words from the dataset, and verified that it is statistically significant by measuring the p-values of the different statistical coefficients, which were all approximately 0.

A.9 Training Hyperparameters and Additional Information

SBA. We train the linear layer of the SBA using gradient accumulation with an effective batch-size of 128, learning rate of 2e-3 with cosine scheduler and a warm-up ratio of 0.03, and train for a single epoch over a single Nvidia-A6000 GPU. All other hyperparameters are set to the default of HuggingFace Trainer.

VBA For the image generation in VBA, we use guidance scale of 9 and 20 inference steps. When generating captions, we decode using beam search with 5 beams.

Distillations. For the distillation, we use batch size of 128, learning rate of 1e-4 with a cosine scheduler, and a warm-up ratio of 0.03 for 2 epochs using a single Nvidia-A6000 GPU. All other hyperparameters are set to the default of HuggingFace Trainer.

A.10 Model Checkpoints Used

We detail here all the checkpoints that were used in our experiments. All model checkpoints are taken from the Hugging Face Model Hub[1]. For the SBA, we used:

- `openai/clip-vit-large-patch14` (only the text encoder)
- `meta-llama/Llama-2-7b`

For the VBA, we used:

- `stabilityai/stable-diffusion-2`
- `Salesforce/blip2-opt-2.7b`

For the distilled model, we used:

[1] https://www.huggingface.co/models

Sentence Conc.			
Method	ρ	ρ_s	τ
LLM with N=3	0.17	0.16	0.15
LLM with N=5	0.19	0.21	0.19
LLM with N=10	0.25	0.25	0.22
ICC	**0.69**	**0.67**	**0.54**

Table 4: **Concreteness evaluation of captions using an LLM with different prompts.** We report the Pearson ρ, Spearman ρ_s, and Kendall τ correlation coefficients. N denotes the concreteness range of possible scores given in the prompt (range of 1-N).

- `distilroberta-base`

For training a captioning model, we used:

- `microsoft/swin-base-patch4-window7-224-in22k`
- `gpt2`

For training a dual-encoder model, we used:

- `bert-base-uncased`
- `google/vit-base-patch16-224`

Appendix B Additional Experiments and Ablations

B.1 LLM-based Concreteness Score

We experiment with an additional method for quantifying concreteness of caption by prompting a Large Language Model (LLaMa-70B-chat Touvron et al., 2023). In order to probe a zero-shot LLM to provide concreteness scores, we used a prompt of the following form:

"`You are a visual expert and you need to provide visual scores for captions according to how concrete they are. You answer only using a single integer number on a scale of 1-N when 1 means the caption is highly abstract and N is a highly concrete caption.`
`Input caption: '⟨caption⟩'`
`Concretenss score is "`

We ablate over three different values of N and report the values and corresponding correlations in Table 4. As illustrated in the table, our method significantly outperforms LLM-based prompting.

We use greedy decoding for all prompts.

B.2 Ablation over the Intermediate Scores

We further verify the importance of using both scores by ablating the effect of filtering with each score in isolation compared to filtering with them combined (*ICC*) on downstream captioning model training. We show the results in Table 5. These results verify that our combined *ICC* score outperforms each score used in isolation.

Data	B@4	M	R	C	S	BSc
LA+SBA	4.4	8.5	20.6	13.0	46	-0.5
LA+VBA	6.4	11.5	27.6	21.5	70	0.31
LA+*ICC*	**6.8**	**12.0**	**28.6**	**24.2**	**75**	**0.32**

Table 5: **Score Ablations** We ablate the importance of using both scores obtained from the two pipelines, over 1M samples of LAION (LA) with similar settings to captioning model training in Table 6.

We also visually show examples of each of the scores' weaknesses and the way they compliment each other. In Figure 5, we show examples of *concrete* captions, the reconstructed captions by VBA and SBA, and the different scores of each of them. The first four rows exemplify why VBA may fail to reconstruct some concrete captions. For instance, the caption "a nurse mopping a surgeon's brow during an operation in an operation pub" was reconstructed to "two people in protective gear" which bears relatively low semantic similarity to the original caption. The main reason these cases happen is due to the inherent difficulty of reconstructing (through a captioning model) from an image the exact caption from which the image was generated, as there may be many possible such captions. In this case, the use of SBA helps determining that the caption is concrete.

In a complementary way, we show in Figure 6 examples of *abstract* captions. In this figure, the first four rows demonstrate that using SBA alone is also not enough, as it is sometimes able to reconstruct abstract captions due to the higher semantic information that is contained in the CLIP embeddings. In this scenario, VBA covers up for these failures, as it is very unlikely to reconstruct abstract text.

These qualitative examples further illustrate the benefit of using both VBA and SBA. Indeed, in both Figure 5 and 6, it can be observed that *ICC* learns to take the best of both worlds, generating low scores for abstract captions, and high scores to concrete ones in a consistent manner.

		COCO Captioning						**COCO**			**Flickr**		
Data	**# samples**	B@4	M	R	C	S	BSc	R@1	R@5	R@10	R@1	R@5	R@10
LA	100k	0.8	4.2	11.1	3.6	18	-0.95	1.7	6.3	10.4	3.0	9.9	16.7
LA+CLIP	100k	0	2.7	7.6	0	2	-0.32	0.2	1.0	1.8	0.5	2.1	3.9
LA+CA	100k	0.4	7.4	18.2	0.9	18	0.18	2.0	7.9	13.2	4.8	15.7	25.4
LA+*ICC*	100k	**5.1**	**11.3**	**31.8**	**9.7**	**45**	**0.39**	**5.0**	**15.9**	**24.4**	**13.1**	**34.6**	**47.2**
LA	500k	0.5	4.8	12.9	1.9	14	-0.04	4.5	15.0	23.0	9.6	27.7	40.5
LA+CLIP	500k	0.2	4.0	11.0	1.1	9	-0.07	2.2	8.0	13.1	4.9	15.1	23.1
LA+CA	500k	0.2	3.9	10.7	1.0	9	-0.07	6.5	19.7	29.3	16.3	40.5	55.2
LA+*ICC*	500k	**7.8**	**12.2**	**30.5**	**21.2**	**72**	**0.35**	**10.0**	**27.1**	**38.4**	**21.7**	**49.6**	**62.4**
LA	1M	0.8	4.2	11.1	3.6	18	-0.95	6.8	19.9	29.2	14.0	38.1	50.6
LA+CLIP	1M	1.0	5.4	12.7	2.8	23	-0.47	5.0	15.3	23.2	9.9	29.0	41.2
LA+CA	1M	0.5	2.5	4.9	1.8	9	-3.9	9.2	25.2	36.0	20.9	49.8	63.4
LA+*ICC*	1M	**6.8**	**12.0**	**28.6**	**24.2**	**75**	**0.32**	**12.2**	**31.3**	**42.8**	**26.4**	**55.5**	**67.5**

Table 6: **Ablation over different dataset sizes**. We perform evaluation over MS-COCO dataset for captioning as well as text-to-image retrieval over MS-COCO and Flickr for different filtering schemes with varying dataset sizes. *Data* denotes the training dataset; *LA* indicates LAION-400M. We compare our performance (+*ICC*) to two filtering baselines; +CLIP indicates filtering by top CLIPScore and +CA indicates Complexity and Action filtering. We also report performance obtained by randomly selecting 100k, 500k and 1M samples. B@4, M, R, C, S and BSc denote BLEU-4, METEOR, Rouge-L, CIDEr, SPICE, and BERTScore metrics respectively, evaluated on MS-COCO Karpathy test split. Best results are in **bold**.

B.3 Ablation over Dataset Sizes

In Table 6, we provide ablations over different dataset sizes for both captioning and representation learning tasks. As is seen there, *ICC*-based filtering outperforms competing methods over 100k, 500k and 1M training samples, further demonstrating the robustness of our method.

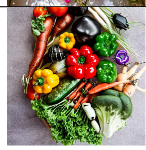
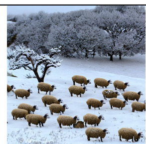
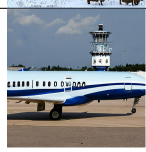
Input caption	SBA reconstructed caption	VBA reconstructed caption	VBA bottleneck image	SBA	VBA	*ICC*
a nurse mopping a surgeon's brow during an operation in an operation pub	a nurse wiping the brow of a surgeon during an operation in an operating room	two people in protective gear		0.77	0.25	0.72
bougainvillea climbing up the wall of a villa	bougainvillea climbing on a wall of a villa	a house covered in pink flowers		0.72	0.26	0.81
table top shot of many vegetables and mexican bugs on a table	close up shot of vegetables and bugs on a table	vegetables arranged in the shape of a human head		0.70	0.25	0.76
silhouette of a man with a gun in poses royalty	silhouette of a man holding a gun in poses royalty	a group of people silhouettes on a white background		0.82	0.26	0.93
small flock of sheep in winter snow on a hilltop	small flock of sheep in snow on a hill	a herd of sheep in the snow		0.72	0.95	1.0
small blue and white airplane parked on the ramp with a control tower in the distance	small blue and white airplane parked on the tarmac next to a control tower	a blue and white airplane parked on the tarmac		0.96	0.95	1.0
a young girl runs through a field of cabbages	a young girl runs through a field of cabbages	a girl walking through a field of cabbage		0.96	0.95	1.0
a red post box and a telephone box stand together in a village	a red telephone box and a post box stand together in a village	a red post box next to a stone wall		0.84	0.89	0.92

Figure 5: **Qualitative Examples for Highly Concrete Captions**. We demonstrate reconstructions of highly concrete captions and the final distilled *ICC* scores. We mark by red low reconstruction scores which correspond to unsuccesfull detection of the concrete captions. As illustrated above, VBA yields generally less consistent scores for concrete captions (see the text for further discussion). Nonetheless, our final distilled scores correctly identify these captions as concrete ones, obtaining high *ICC* scores over these captions.

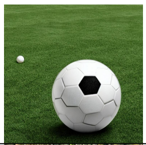
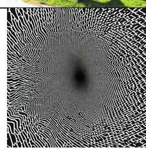
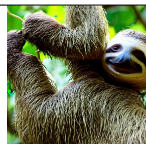
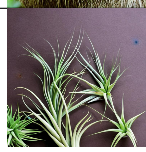
Input caption	**SBA reconstructed caption**	**VBA reconstructed caption**	**VBA bottleneck image**	**SBA**	**VBA**	***ICC***
keep an eye on the ball when it comes to investments	keep an eye on the ball when it comes to investments	a soccer ball on a green field		0.91	0.19	0.1
what 's the best thing about having a best friend of the opposite gender ?	the best thing about having a friend of the opposite gender	two young women sitting on a bench		0.89	0.16	0.1
film character : would you like to bet on these shares this christmas ?	which film character would you like to see in your shares this christmas?	santa claus, santa claus and sant		0.79	0.1	0
this is located in my home town !	this is located in my hometown!	a sign in front of a statue		0.75	0.28	0
chaotic systems are sometimes described using fractal patterns	fractals are patterns that can be found in many forms, such as chaotic systems and natural structures.	a black and white tunnel		0.22	0.19	0
on an average , the sloth travels feet a day	a sloth spends most of the day on its feet	a sloth hanging from a branch		0.17	0.27	0
get tips for biological genus , more commonly known as air plants , in your home	learn how to care for air plants, one of	a bunch of air plants on a brown surface		0.32	0.25	0
versatile and highly capable , there 's more to this tiny camera than its giant zoom	this little camera packs a big punch with its zoom lens and 2	a camera on a wooden table		0.25	0.24	0

Figure 6: **Qualitative Examples for Highly Abstract Captions**. We demonstrate reconstructions of highly abstract captions and the final distilled *ICC* scores. We mark by red captions which were reconstructed well (note that in the case of abstract captions, high scores correspond to unsuccessful detections of the abstract captions). As illustrated above, SBA yields generally less consistent scores for abstract captions (see the text for further discussion). Nonetheless, our final distilled scores correctly identify these captions as abstract ones, obtaining low *ICC* scores over these captions.

a bundt wedding cake with white chocolate dripping, evergreens, pinecones and sugar powder to imitate snow | A young boy stands in front of a wall with height measurement marks and has his hand up to show | Elderly man with a cocktail during holidays | silver soda can and glass with white background - Stock Photo | foto of wallabies - Portrait of a wallaby in the nature - JPG | Brindle pendant in light grey with copper interior | Stock photo of homemade cookies and a cup of coffe | Girl in the gym lifting up the barbell - silhouette | Young blond woman wearing a dress in the forest | A fish eye photo of people climbing high ropes | Young alligators basking in the sunlight | Couple sat by basket full of grapes | vintage book and light bulb on wood table | tree with birds and birdcages vector image | A boy volunteer with birds on his shoulders | Colorful streamers hanging from the ceiling. | A glass bowl full of yellow cream and red and orange coloured fruit on pink & white background | A Chinese man walks past a billboard for a new commercial development which reads 'Shangrila is in your mind but | Shopping bags isolated on the white background | child tying his own shoes | A spotted harrier cruising to look for food | Young girl in a party dress looking bored and unhappy | Avatars of a male and in business suits. | a stack of miso chocolate chip cookies on a white plate | Bobcat in a hollow log

foto of florida-orange - Jacksonville skyline in orange background in editable vector file - JPG | Old man cleans tables at KFC to take home leftover food for family | c1913 to 1955 tall stack of music, opera and ballet books sg | Tropical Leaf Necklace 16 | welcome to india card with famous landmarks vector image | Set of empty picture frames for your own vector | castle hotel and spa wedding photos, ceremoy and reception | Radar monitor - Aircraft radar for airport with world map... | Grey and yellow consulting or planning concept infographics set | A darkened hall filled with server racks on either side and a silhouette of a Facebook worker at the end, | 100Ducati Desmoquattro at 2009 Seattle International Motorcycle Show 2.jpg | A film still shows two panels, with green ink. On one of them, the letters RELIC can just be made | spence cabin weddings | businessman in modern office writing ghostwriter in the air | large patio roof with adjustable louvres for outdoor seating weather protection and shading | pic of gesture - vector illustration of collection of hand gestures - JPG | chubby woman eating on scale stock photo, chubby woman on the scale eating a yogurt, isolate on white by iMarin | Fisherman on a small blue and white boat with fishing net between the waves of the sea. Liguria Italy | pictures of bedroom architectural details from hgtv | EHM water ionizer and alkaline water machine factory on sale | Christopher Boffoli's photograph of a toy motorcycle rider, jumping over three toy cars and a slice of cheesecake | Pirate kids and their treasure | Incotex Benson Straight Leg Wool Trousers | QR Code for Florida Virtual School at local Shell Gas Station | Peace Love Colorectal Surgery Oval Sticker (50 pk) | moonstruck chocolates | creative eyelashes - closeup of the eye of woman with... | Simply Perfect Braided Wedges | New Years Fireworks in Seattle, 2011->2012 | 2008 Volkswagen GTI Photo | A sport fishing boat heading out of Wanchese harbor of the Outer Banks at dawn for a day of off-shore | A young beautiful girl holding a wild fox animal that was traumatized by a man and rescued by her and

Zero Zebra Safari Party Dairy-Free Chocolate Animals | The rescued soccer team members pose with a sketch of the Thai Navy SEAL diver who died while trying to | How to get a bobcat out of your window blinds | Pocket watch: technically interesting pocket watch with rare crown winding in manner of O. | I am an Aspie Girl A Book for Young Girls with Autism Spectrum Conditions by Danuta Bulhak-Paterson, Tony Attwood | Trump on etch a sketch | floor plan drawing software create your own home design easily | 1000 ideas about lake house plans on pinterest house for Basement planner online | Chanel 5, the first perfume i received as a gift. Love it! | lush greenery at National gallery of modern arts - Bangalore | A lounge room of greys and creams, black and white prints all come together to make this a relaxing and | christmas bible verses for preschoolers five scriptures about children should 664 | MAJESTIC PET PRODUCTS - Santorini Chevron Round Pet Bed - pet bed looks great in any room of your house | Sanusi under house arrest, moved to a 2-bedroom apartment without electricity & access roads (Photos | Novak Djokovic (Ser) defeated Juan Martin Del Potro (Arg) in US Open final
 Flushing Meadows 09-09-2018 US Open
 | 2013 men's the novelty original t-shirt with patterns Double-headed eagle and RUSSIA sizel xl xxl xxxl 4xl shirts free shipping | Poster of Seven Below | EFCC operatives evacuating the safe, the house where the cash was hidden and the money | closet converted into mudroom | make a closet more functional by removing doors, adding a bench at kid height, hooks | How to install an SSD in a laptop | computer tutorial | The logo of German carmaker BMW is seen on a car displayed during the annual results press conference in Munich, | Well Established and Well Equipped Butchers, Fruit and Vegetables, Frozen Seafood Plus Convenience Groceries, South London for sale | BMW Is Looking Into Gas-Powered Vehicles | summer fashion scarvesnew scarf trendswhite by scarvesCHIC on Etsy, $15.90 | Celebrating 20 years, EGEC shares declaration on the great role of geothermal energy | capricorn tattoos designs ideas and meaning tattoos for you | 25 best ideas about spice storage on spice | Julbo Eyewear - Atmo Goggle (4-8 Years Old) (Red Trans Orange Lens) Snow Goggles | The Roman temple of Jupiter is seen in the background as Lebanese youths play in the snow on January 9, | The DJI Osmo kit includes the grip, gimbal, camera and device holder for your smartphone (there is a companion app). | Miyake celebrates with his team after winning a silver medal at London 2012. | National MS Society's Katie Boothroyd, Board Member Joan Ohayon. Photo by Tony Powell. Tea Honoring Women of the Diplomatic Corps. | Top sweet and fortified wines of 2012 | there is also a small laundry with all-in-one washer and dryer | change my background how to change desktop background in windows 10 | Hand wrapping Basics - How to wrap your hands for boxing, kickboxing, and Muay Thai with long wraps | favorite colors on taupe benjamin and paint colors | Black & White Houndstooth Infinity Pocket Scarf - Travel Scarf - The Poppy Stock | looking out to the yew garden | An entrance door to transform your home for Home front entry doors | Image result for paytm with modi advertisement | Drawing notebook. never thought of including this inside a felt book, always had a separate art bag.. | how to remove a kitchen tile backsplash | ONLY - Cara Long Sleeve Shirt (Navy Blazer) Women | Tying an olive dun with mallard wings | All cold and hot rolling seamless steel pipe diameter | and when the clock strikes midnight halloween will be over | Another member said she'd tried for years to get her floor to look this clean, declaring she 'loved' the bargain | 2PCS/Set Hematite Natural Stone Bracelet1 | architectural drawings for sale unknown vintage architectural drawing for sale at 1stdibs | Prince Harry to lose honorary military titles as palace confirms exit date from royal family | Explore One 70mm Telescope | The mess I left at the end of the school year in 2015 when I couldn't unpack or move into | A pretty paper garland adds a festive touch. | Hydroponic Fodder ProFeed Growing System | Antioxidant rich RED salad with lemony dressing is delicious lunch! | Sunflowers on Saturday, when I felt called to ask David to photograph me with them since they played such a | pencil crafts for back to school and beyond | COLROVIE Culotte Leg Elegant Cami Jumpsuit Women Box Pleated Sexy V Neck Jumpsuits 2017 Fall Surplice | how to make fabric flower rosettes, tutorial | new cars with best warranty all about extended auto warranty contracts leadhub | 30 sliding barn door designs and ideas for the home With barn door wide opening | Foreign stocks for students and grannies | how to wear flat shoes to a wedding | A cartoon on the situation with languages in Ukraine cartoon. | Modest partisan differences in views of elected officials | interior design for my home minimalist interior design is maximum on style | Ebook download and read online electronic book button or icon | what do you want to know about alta motors electric street tracker | Raw Oysters are the perfect food to increase your testosterone! | Minecraft cake - Both tiers are vanilla cake with vanilla buttercream covered in fondant. The tiles are all modeling chocolate | cbd infographic why patients are leaving big pharma | the majestic elephant - one of the big 5 |

Car insurance advice: How to keep your car safe in winter weather | We only supply the tire. If there is a rim shown in the picture, it is for display purposes only. | "Steven Wright Quote: ""It's like the Wild West, the Internet. There are no rules.""" | pathandpuddle: How long do animals live? | A lack of sleep could be caused the nutrients in your diet | No time to explain. Just put on the hats and act casual. | His last at-bat, a pop fly to center field. #garrettreade #littleleague #thatsmyboy | Higher consumption of sugary beverages linked with increased risk of mortality | Words of Gymnastics Terminology w/ Monogram Drawstring Bag | Zaanse Schans, Netherlands - May 5, 2015: Tourist Visit Windmills And Rural Houses In Zaanse Schans | Losing out: BP will temporarily be locked out of lucrative deals, including contracts to supply the US military with fuel. | QuickBooks - Access | "Cranberry Chevron Rug - Deep red hues cut a rug here. The chevron is a ""go with anything"" pattern and | Bomag reports that their single direction vibratory plate compactors are great tools for contractors' day-to-day use in soil and asphalt | "Augustabernard bias-cut satin evening gown, c.1930. Label: ""Augustabernard"" with a stamped couture number on the back." | How to graduate as a successful edupreneur | garland for stairs christmas house tour decorating ideas how decorate for | This Mexican Layer Dip is easy to make and full of flavor! With layers of spicy black bean dip, homemade | what is a research process paper The term research paper may also refer to a scholarly article that contains the | Ammonia is often used in cleaning products because it reacts with grease making this easier to remove | They do not take ownership of valuable deposited with them? | Can i push out my wall to get an 8x8 bathroom leave me for Small bathroom design 5 x 8 | Making one of these wall hangings is a great way to use up old yarn ... | 6 Pack - 16 Ounce Grolsch Bottles with Easy Cap Flip Top Caps for Brewing Beer, Kombucha, Kefir, Water, Thick | Annual: The event, celebrated every year to herald good monsoon rains for increased rice harvest, prosperity and goodluck, is one | Graphic on Australia's Tasmanian Devils, rare carnivorous marsupials in a battle for survival against a contagious facial cancer. It's been | remodeling small bathroom ideas on a budget small bathroom remodel on a budget brown ceramic tile floor walk | When a 17-year-old Sharapova burst onto the scene in 2004 by winning Wimbledon, she was an immediate hit. Sure, her | The course has a convenient location in the community of Fernie. | Famous burj al arab hotel dubai 6 said to be the for Dubai world famous hotel | Among Us. (Innersloth) | Download car and vehicles decal graphics kit designs ready to print and cut for vinyl | GSM Functionality GSM Technology is a special design which can be used in conjunction with a variety of signallers depending | Kendal Town Council are calling for a bypass to solve traffic problems after Storm Desmond | Are You Am I - Faira Dress **white | interior home design also with a interior wall design also with a | One reason Millennials book cruises is the low cost - contrary to popular belief | Containers have grabbed a large share of the intermodal freight business, and here's a miniature trainload of them at the | be a donor be a hero | "Operation Surf Santa Cruz is an annual event that honors active duty military soldiers through ""an epic life-changing surfing experience.""Many | hyster h40 h forklift will not go forward or backward stuck brakes | Nominated banks roles and responsibilities under a letter of credit transaction. | Kingsbrook animal hospitals blog preferred veterinary care in two beloved basenjis kylie and cricket in a house fire in april | Example of TRAM:Cross-Domain-Solution | Skip the recycling. Use your soda cans to make bracelet cuffs instead! | They rise in quite an interesting way as well. Keep an eye on them so they don't burn, and make | "15 Thanksgiving Day Ideas for Couples. Holidays get so busy we sometimes forget our priorities. Keep your spouse at the | Tao Xiangli gets out of his homemade submarine after operating it in a lake on the outskirts of Beijing September | Professionally edit your voiceover, audiobook or podcast | "Eli struggling to find cell service ""under the biggest cell phone tower in Paris""" | Don't have time to shampoo your hair, but still want to look glam? Me too! And that's why I love | Vivid Vision for Success | Rich Karlgaard | The pistol 01 concept is a very precise and reliable sidearm! | buy a domain name | Tate Stevens - Winner of 2012 X Factor, Simon Cowell Stock Photo | Click this cover for a(n) eBook sample of Choke Point | FI Week #5 #6 #7 : Let the new digital influencers shine | Travertine is a perfect partner2 | STF has arrested the wanted accused who recruited fake teachers | SUMMER HITS 2016 Mixed by DJ Golan | Springet fra Big Business til Smaller Business (Starting Over) | Banque De Francs, French Banknote Assortment, ca. 1945-1961. | tarek christina tarek and christina el moussa s divorce affects their | Phoenix Suns are changing their perception to the basketball world | Tan Cartoon Doctor Man Carrying His - African American... | 2013 New arribal digital holly quran mp4 player wholesale price and 28 language translation(China (Mainland)) | Your spa decor should reflect your target market's preferences. | hwaseong-fortress-suwon-part-2 | Salman and Sonakshi in Dabangg 2 |

Figure 7: **Manually Annotated Captions**. The captions are sorted according to concreteness, where captions with thehighest score illustrated in the top cluster and lowest at the bottom cluster. We truncate captions that are longer than 20 words, and separate captions by |.