

ICC: Quantifying Image Caption Concreteness for Multimodal Dataset Curation

Moran Yanuka Morris Alper Hadar Averbuch-Elor Raja Giryes

Tel-Aviv University

<https://moranyanuka.github.io/icc/>

Abstract

Web-scale training on paired text-image data is becoming increasingly central to 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 or subjective text. These approaches lack the fine-grained ability to isolate the *most concrete* samples that provide the strongest signal for learning in a noisy dataset. 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 unsupervised 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 caption-level texts. Moreover, we show that curation using *ICC* complements existing approaches: It succeeds in selecting the highest quality samples from multimodal web-scale datasets to allow for efficient training in resource-constrained settings.

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.

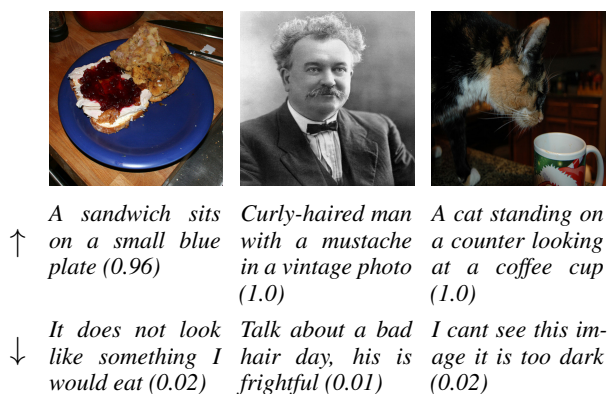


Figure 1: Given an image caption, *ICC* measures its visual concreteness. We show samples from MS-COCO (Lin et al., 2014) illustrating captions annotated by different annotators with low (\downarrow) and high (\uparrow) *ICC* scores. As seen above, our method successfully differentiates between concrete and abstract or subjective captions, even for high-quality datasets such as MS-COCO. This is done by quantifying visual-semantic consistency using multimodal foundation models.

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), i.e. captions irrelevant to their images; using rule-based proxies such as measuring the complexity of captions via semantic parsing (Radenovic et al., 2023); or removing images that contain text that overlaps with the caption (Maini 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, yielding high CLIP similarity, but contains subjective details which provide a confounding signal when training VLMs (See also Figure 2). A model trained to generate such captions from images may learn to hal-

lucinate details, e.g., liking a certain type of food in our example, which are not visually grounded and are highly subjective. Similarly, such image-caption pairs provide a weaker signal for representation learning than images with visually concrete captions (e.g. “A sandwich sits on a small blue plate”), which may impede the learning process – particularly in a resource-restricted setting where data or compute is limited.

Thus, we suggest filtering image captions by their *visual concreteness*, referring to the extent to which a text describes visual aspects of a scene in a manner that can be vividly imagined (Schwanenflugel, 2013; Hessel et al., 2018)¹. This contrasts with abstract text, which may correspond to many possible visual interpretations or include subjective information. We show that this new dimension of textual quality enables selecting image-caption pairs that provide a strong supervision signal for vision-and-language tasks, particularly in resource-constrained settings where training directly on noisy web-scale multimodal data fails to converge to a satisfactory solution in a limited number of iterations.

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 unsupervised autoencoding pipelines with visual-semantic information bottlenecks. Specifically, we use a visual-bottleneck autoencoder that leverages text-to-image generative models’ competence and a semantic-bottleneck autoencoder that identifies how well a large language model (LLM) recovers the input caption from its semantic CLIP embedding. As these models require costly inference through large generative models, they cannot feasibly run on a large scale; therefore, our *ICC* metric is distilled from these pipelines, enabling fast, computationally-efficient inference.

In our experiments, we demonstrate that when dealing with limited training iterations, employing *ICC* for filtering multimodal datasets leads to enhanced performance in image captioning and representation learning. Moreover, our results indicate a strong correlation between *ICC* and both

¹Some works have treated this as roughly synonymous with *imageability* (visual association), while others use *concreteness* to refer more generally to association with sensory experiences of all types (Richardson, 1975; Khanna and Cortese, 2021). Our work focuses on the visual modality.

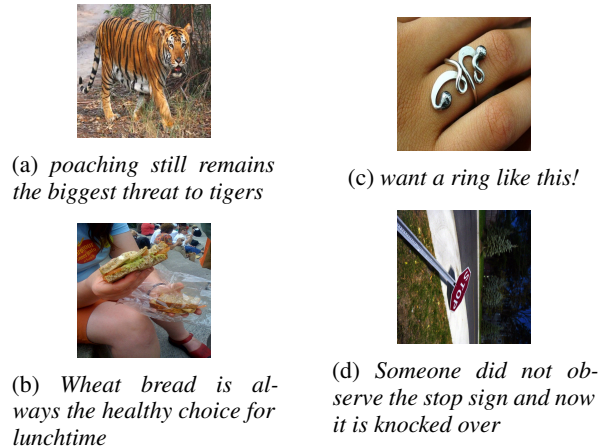


Figure 2: **Examples with high CLIP similarity and low ICC.** We show examples from Conceptual Captions dataset (a) and (c), and COCO dataset, (b) and (d). While these captions are semantically related to the images, they are abstract or contain subjective non-visual information that, unlike *ICC*, CLIP fails to detect.

single-word concreteness and caption text scores.

Stated explicitly, our contributions are as follows: (1) We propose the *ICC* metric distilled from foundation VLM models with a novel combination of unsupervised autoencoding pipelines; (2) we show that *ICC* highly correlates to human concreteness judgements of caption texts; (3) we demonstrate that *ICC* succeeds in selecting a core of samples from web-scale image-caption datasets for vision-and-language tasks, with superior downstream performance to existing filtering methods.

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 a visual representation with less information loss. Conversely, we expect that visually abstract or subjective text cannot be converted to or from a visual representation without significant information loss, since it does not clearly describe a well-defined image.

As an example, consider the text “Wheat bread is always the healthy choice for lunchtime” in Figure 2b. The notion of wheat bread being a healthy choice is inherently non-visual and is unlikely to be directly depicted in an image. Therefore, this information is likely to be lost in an autoencoding process that includes an image as the bottleneck, when the encoded image is decoded back to the textual modality.

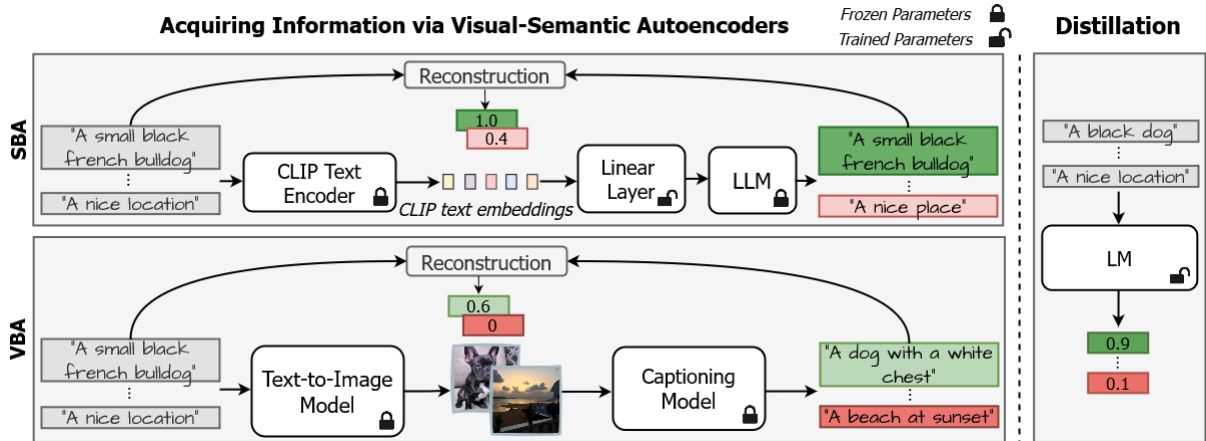


Figure 3: **ICC pipeline for predicting visual concreteness of image captions.** We first acquire training data using a semantic-bottleneck autoencoder (SBA, top left) and an visual-bottleneck autoencoder (VBA, bottom left). We then distill a weighted combination of their reconstruction scores into a smaller language model (LM, right), which learns to produce *ICC* scores for new text. We visualize reconstruction scores for highly concrete (“*A black dog*”) and highly abstract (“*A nice location*”) text. High and low scores are colored in green and red, respectively. Our final score, which combines the two pipelines, yields more accurate concreteness predictions than each of them.

We model this effect with multimodal autoencoders (Kamath et al., 2023; Yang et al., 2023). In our setting, we use multiple autoencoder components that convert text to and from visual-semantic representations using foundation VLMs, and quantify the information loss of this process as a proxy for visual concreteness. While these autoencoders provide a strong signal, they are composed of slow, computationally-intensive large generative models making inference infeasible on a large scale. Therefore, we distill their scores into a small model which allows for an efficient calculation of the *ICC* scores.

We proceed to describe our proposed visual-bottleneck autoencoder and semantic-bottleneck autoencoder components, and their distillation into the final *ICC* metric. See Figure 3 for an overview of our full pipeline.

Visual-Bottleneck Autoencoder (VBA). Since a caption represents an image, we construct the VBA by using an image as an intermediate representation via which textual information passes. In particular, we concatenate a frozen text-to-image model (Stable Diffusion 2, Ramesh et al., 2022) and a frozen captioning model (BLIP-2, Li et al., 2023) as shown in Figure 3 (bottom left). This autoencoding pipeline measures text concreteness by encoding and decoding a caption, followed by measuring semantic fidelity in reconstruction using BERTScore (F1) (Zhang et al.,

2019). We note that this pipeline contains *no trained parameters* as it concatenates pretrained, frozen models.

While the VBA pipeline is a simple and intuitive way of enforcing a visual bottleneck, it may sometimes produce sub-optimal reconstructions even for highly visual texts due to its inherently lossy nature. For example, the caption “*a small black french bulldog*” in Figure 3 may be reconstructed by the VBA from the generated image to “*a dog with a white chest*”, which is relatively semantically far from the original caption and thus results in a relatively low reconstruction score of 0.6 for a concrete caption. This stems from the dense information content of generated images, which may contain details (such as the dog’s white chest) which were not mentioned explicitly in the original caption, and from the tendency of the captioning decoder to focus on different details than those used to generate the image. To alleviate this issue, we proceed to propose a complementary method using a stronger prior on caption semantics.

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 3 (top left). We extract visual information from the CLIP text embedding space by utilizing a frozen

LLM (Llama-2-7b, Touvron et al., 2023), by training a linear layer that converts CLIP 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. By keeping the encoder backbone (CLIP) frozen, this introduces an information bottleneck preventing faithful reconstruction of abstract texts.

After training the SBA over image–caption pairs, we use it for measuring text concreteness by encoding and decoding the text followed by measuring reconstruction fidelity. To measure preservation of fine-grained textual details, we quantify this fidelity via per-character edit distance (Levenshtein et al., 1966), standardized by caption length, as detailed in Appendix A.1.

This pipeline generally succeeds in reconstructing highly concrete text (such as “A *small black french bulldog*” shown in the top left part of Figure 3). However, the strong textual prior of the SBA may also leak information about abstract and subjective captions as well (e.g. the abstract caption “A *nice location*” yields a relatively high reconstruction score of 0.4), limiting its correlation with visual concreteness. Overall, the SBA and VBA provide complementary scores, where each correlates more strongly to visual concreteness in different cases. Therefore, they perform most strongly when combined together, as we explicitly verify in our ablations in Section 4. We also show qualitative examples in figures 8 and 9 in the appendix.

ICC Distillation. Using the aforementioned pipelines to quantify the concreteness at scale is not feasible, as this requires running large models (e.g., diffusion models, LLMs) with billions of parameters for many forward passes per instance (up to dozens of forward passes for the diffusion models inference and for the LLM and captioning model decoding). This requires more than 1,000 GPU hours for a dataset of 1M samples. Therefore, we assemble SBA and VBA reconstruction scores over a relatively small collection of image-caption pairs and distill their aggregated values into our final ICC score. This enables efficient inference that can easily run on a large scale, with over a hundred times faster inference time and much less compute required. Specifically, we train a small text encoder model (Liu et al., 2019) to predict a logit-linear combination of the SBA and VBA scores, computed as described in the appendix.

Implementation Details of ICC Construction.

For the construction of our ICC score, we use a subset of CC3M (Sharma et al., 2018) composed of 595K image-caption pairs, introduced by Liu et al. (2023) and designed to have wider concept coverage. We take a subset of 476K samples for training the linear layer of the SBA, and train for 2 epochs with a batch size of 128 and learning rate of $2e-3$ with cosine scheduling function. The remaining 118K samples are used for generating reconstruction scores through the VBA and the trained SBA. For each input caption, we generate five reconstructed captions using beam search (five beams) with the VBA’s captioner and the SBA’s LLM and then choose the reconstructed caption with the highest similarity to the source caption. By generating the reconstructions and measuring the reconstruction fidelities, we obtain a dataset of 118K captions and corresponding reconstruction scores. We standardize by caption length to disentangle the dependency of the reconstruction scores to the caption length (i.e., forcing the same distribution of scores for all caption lengths), as described in the appendix. We train a small language model (DistillRoberta-Base) to predict the combined scores on these samples with a Mean Squared Error objective. This final distilled model is used for generating the ICC scores.

3 Results

We turn to show ICC’s benefit in data curation for downstream tasks (Section 3.1), followed by its correlation to human judgement (Section 3.2).

3.1 VLM Dataset Curation

Experimental Settings. We investigate the effect of ICC and other filtering methods for curating a core of high-quality image-caption pairs from large multimodal datasets, comparing their effects on downstream task performance – both discriminative (representation learning) and generative (image captioning). We follow similar settings as described in the Datacomp (Gadre et al., 2023) benchmark’s filtering track², with the following modifications to model the resource-limited setting: given a training dataset comprised of \mathcal{M} samples, the downstream model is constrained to train for exactly $\mathcal{N} \ll \mathcal{M}$ iterations over the filtered subset of the dataset. This contrasts with

²As opposed to the BYOD track which allows for modifying the samples, for instance by using synthetic captions.

the original Datacomp setting where $\mathcal{N} = \mathcal{M}$, which requires significant compute for a web-scale dataset. Our formulation tests the ability of filtering methods to curate high-quality core subsets of such datasets. Our initial subset of LAION-400M is composed of $\mathcal{M} = 8M$ samples and we fix $\mathcal{N} = 2M$ training iterations. To verify the robustness of our method, we measure downstream performance over visually grounded benchmarks across three different sizes of filtering.

We compare to four existing filtering methods – CLIPScore (Hessel et al., 2021), Complexity and Action (CA) (Radenovic et al., 2023)³, T-MARS (Maini et al., 2023), and PACScore (Sarto et al., 2023). CA is a rule-based filtering method which aims to retain only sufficiently complex captions that also contain an action, based on semantic parsing. T-MARS filters multimodal datasets by removing samples whenever an image includes text that overlaps significantly with the caption. PACScore trains a CLIP-based model with positive-augmented contrastive learning approach, showing improved correlations with human intuition in scoring image-caption pairs. As opposed to these methods, we focus on filtering according to the concreteness of image captions.

Captioning Models. In Table 1 we show quantitative results of applying *ICC* filtering on top of standard CLIPScore filtering over the subset of LAION-400M for training a captioning model. 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. We use a batch size of 100, and learning rate of $2e-5$ with a cosine scheduler. We test our approach over two standard captioning benchmarks datasets – MS-COCO (Lin et al., 2014) and NoCaps (Agrawal et al., 2019), across multiple captioning metrics (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004; Vedantam et al., 2015; Anderson et al., 2016; Zhang et al., 2019; Wada et al., 2024). As illustrated in the table, filtering with *ICC* outperforms by a large margin the alternative filtering methods for captioning given a fixed number of desired samples and training iterations. Note that unlike other methods, *ICC* is directly aligned with the captioning objective, as a captioning model should generate visually-grounded concrete text. This may explain the large

³Using our re-implementation, as there is no publicly available code.

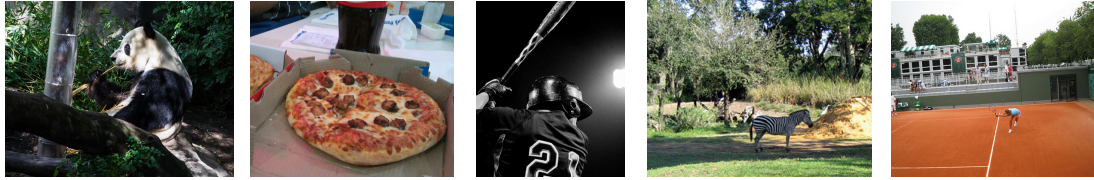
gap in performance between *ICC* and other filtering baselines. We show qualitative comparison between captioning models trained with different filtering methods in Figure 4, exemplifying how filtering with *ICC* promotes more concrete and accurate captioning.

Image-Text Representation Learning. We also perform a representation learning experiment by training a dual text and image encoder model on LAION-400M filtered with different methods. Table 2 reports text-to-image retrieval over standard held-out retrieval benchmarks, namely MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015). The model is initialized from pretrained vision and text encoders (ViT-base, BERT-Base) (Dosovitskiy et al., 2010; Devlin et al., 2018), as suggested by Zhai et al. (2022). We use a batch size of 128, learning rate of $2e-5$ with a cosine scheduling function. All other filtering methods in the table are identical to the ones in the captioning setting. As illustrated in the table, *ICC* yields superior performance for this task, showing that our method selects samples which provide better signals for downstream retrieval applications in this setting.

We note that although prior work has found filtering methods such as CLIPScore (Hessel et al., 2021) to be beneficial (Gadre et al., 2023), we find that it fails to significantly improve (or even degrades) results in the case of selecting a small core of samples. This accords with previous work showing that applying filtering to LAION-400M with CLIP degrades the performance (Maini et al., 2023) in some of the benchmarks, likely due to high-scoring images containing literal text that overlaps with the caption.

3.2 Concreteness Correlation

Table 3 shows the correlations of different concreteness estimation methods to ground-truth concreteness scores on both single-word and caption-level benchmarks. We compare *ICC* to three baselines. The first baseline is zero-shot probing of CLIP through Stroop probing (SP) as proposed by Alper et al. (2023). The second baseline is aveCLIP (Wu and Smith, 2023), a learned metric quantifying concreteness at the sentence level, which generates multiple images from a caption and measures the average CLIP-similarity between the text and generated images. Due to its high computational cost, we evaluate it on a ran-



CA	<i>Tiger cubs playing in the rain at the Zoo of the Ozarks in Washington, D.C. on Saturday, Oct. 18</i>	<i>Coffee at the bar. I love this place! It's a great way to get away from the hustle and bustle</i>	<i>Cleveland Indians vs. Boston Red Sox</i>	<i>Cambodia, the largest of all the African savannas, is one of the most arid regions in the world.</i>	<i>Rugby World Cup 2019: The men's singles final takes place at the Ritz-Carlton in London, England, on Saturday,</i>
TMS	<i>Catching a lion in the wild is one of the most beautiful things you can do in the wild.</i>	<i>Coffee at the bar. Photo credit: Flickr userfairy.com.au. (via Flickr)</i>	<i>Buster Posey hits a two-run home run...</i>	<i>Aerial view of a wild boar in a field in Namibia, South Africa, Africa. (Photo courtesy of the Namibian Wildlife)</i>	<i>Astonishingly, there was no shortage of competition between the two-school teams at the London 2012 Olympic Games.</i>
CLIP	<i>Polar bear cubs pose for a photo with a polar bear in the background. Credit: NASA/JPL-Caltech/UCLA</i>	<i>Pizza Hut Creamy Pizza Sandwich with Bacon, Cheese, and Tomato Sauce</i>	<i>Bryce Harper of the Toronto Blue Jays signs autographs for fans prior to the game</i>	<i>Aerial view of the world's largest crocodile in the Serengeti National Park.</i>	<i>New York Mets Fanatics Authentic 8" x 10" Skateboard Deck</i>
ICC	<i>Panda eating bamboo</i>	<i>A picture of a pizza box full of pizzas</i>	<i>black and white photo of a baseball player</i>	<i>Zebra at the zoo</i>	<i>A view of the tennis court from the front.</i>

Figure 4: Qualitative examples of captions generated by captioning models trained on datasets filtered with different filtering methods, over images from MS-COCO test split. CA denotes Complexity and Action filtering, and T-MARS is marked by TMS. As seen above, models trained on ICC-filtered data generate much more concrete and visually-grounded captions.

dom subset of each benchmark (as described in the appendix). Finally, we compare to GPT-3.5-Turbo and GPT-4o (Achiam et al., 2023), used in the zero-shot setting by prompting them to provide concreteness scores. The prompts used are detailed in the appendix.

Correlation to Word Concreteness. We first validate our metric by measuring it on the dataset introduced by Hessel et al. (2018). This consists of 39,954 English unigrams 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 seen in Table 3, ICC outperforms prior dedicated methods for measuring word concreteness, while performing competitively with the proprietary and much larger GPT-4o.

Correlation to Caption Concreteness. We manually annotate concreteness scores for 500 captions from LAION-400M (Schuhmann et al., 2022), selected to cover a wide variety of levels of concreteness. As seen in Table 3, ICC outperforms existing methods in this setting by a large margin, demonstrating its advantage in selecting the most concrete image captions.

4 Ablations

Distillation Concreteness Effect. Although the distillation procedure is necessary to make inference feasible with respect to runtime, we provide further motivation by measuring the effect of distillation on the correlation to ground-truth annotations of concreteness scores in Table 4. As can be seen, the distillation improves correlations values, providing further motivation beyond compu-

MS-COCO									NoCaps						
Method	# Samples	B@4	M	R	C	S	BSc	P	B@4	M	R	C	S	BSc	P
Random	100k	0.9	4.7	11.2	5	2.3	0.64	0.18	1	5.1	11.9	5.6	1.6	0.68	0.11
CLIP	100k	1.1	5.5	11.9	2.5	2.2	0.75	0.12	1.4	5.7	12	3	1.5	0.71	0.08
CA	100k	0.9	3.7	7.3	3.2	1.6	0.27	0.20	1.6	4.4	9.4	4.1	1.2	0.33	0.11
T-MARS	100k	1.2	4.6	10.6	5.6	2.3	0.53	0.20	1.3	4.9	11.6	6.3	1.7	0.61	0.11
PACScore	100k	1.9	6.8	15	5.5	3	0.82	0.16	2.9	7.5	16.1	7.7	2.4	0.79	0.1
<i>ICC</i>	100k	10.1	15.4	35.4	35.8	10.3	0.9	0.39	12.1	15.8	35.9	33.3	6.4	0.9	0.2
Random	200k	0.9	4.2	9.8	5	2.2	0.51	0.19	1	4.8	11.2	5.9	1.7	0.6	0.11
CLIP	200k	1.3	5.7	12.4	3.4	2.6	0.72	0.21	1.6	6	12.6	3.5	1.8	0.67	0.09
CA	200k	0.5	2.8	5.7	3.7	1.3	0.18	0.20	1.2	3.4	7.2	4.1	1.1	0.24	0.12
T-MARS	200k	1.1	4.6	10.7	6.5	2.4	0.5	0.21	1.7	5.4	12.3	7.8	1.9	0.6	0.12
PACScore	200k	2.9	7	15.5	7	3.7	0.73	0.19	3.8	7.4	15.8	8.8	2.6	0.67	0.12
<i>ICC</i>	200k	10	15.2	34.6	35.5	10.4	0.9	0.39	13.1	15.8	35.2	34.3	6.7	0.9	0.21
Random	500k	0.6	3.4	8	4.5	1.9	0.42	0.2	0.9	4.2	10.1	5.5	1.5	0.55	0.12
CLIP	500k	5.2	9.4	22	15.1	5.3	0.8	0.24	5.2	8.9	21.3	12.9	3	0.8	0.13
CA	500k	0.7	3.1	6	3.6	1.4	0.19	0.2	2.1	4.5	9.4	5.3	1.5	0.29	0.13
T-MARS	500k	0.8	3.7	8.9	5.7	2	0.42	0.21	1.2	4.7	10.8	6.5	1.7	0.65	0.12
PACScore	500k	2.6	6.5	15	8.6	3.7	0.65	0.21	3	6.9	15.4	10.4	2.6	0.65	0.13
<i>ICC</i>	500k	8.3	13.9	31.4	30.9	9.7	0.89	0.37	10	14.2	31.3	28.2	6	0.89	0.2

Table 1: **Captioning results for different filtered dataset sizes.** We perform evaluation of captioning models over MS-COCO and NoCaps datasets trained over different filtering schemes of the LAION-400M dataset, with varying dataset sizes. We compare the performance of *ICC* to five filtering baselines. Among these, Random refers to random samples from LAION-400M, CLIP indicates filtering by top CLIPScore, and CA indicates Complexity and Action filtering. B@4, M, R, C, S, BSc and P denote BLEU-4, METEOR, Rouge-L, CIDEr, SPICE, BERTScore and Polos metrics respectively. # Samples denotes the amount of samples retained after filtering. Best results are in **bold**.

tational efficiency and simplifying the inference of our *ICC* model. We hypothesize that this improvement is due to smoothing of noisy reconstruction of the VBA and SBA by the distillation process.

Distillation Speed-up. We ablate the speed-up provided by the distillation phase by running the SBA, VBA and the distilled *ICC* on the same hardware settings (an Nvidia A6000 GPU), the same batch size of 1 and the same caption samples. We find that the SBA and VBA process 0.45 and 0.2 samples per second respectively, and the distilled score processes 45 samples per second. Note that the time it would take to generate scores for our 8M subset of LAION-400M dataset is approximately 11,000 GPU hours for the VBA and 5,000 GPU hours for the SBA compared to just 50 GPU hours using the distilled *ICC*. Additionally, for a batch size of 1, the distilled model takes less than 700 MB of GPU memory compared to 13GB and 14GB for the VBA and SBA respectively.

Use of Both SBA and VBA Scores. We also ablate the use of both SBA and VBA scores for downstream captioning model training in Table 5. In the figure, we show captioning metrics (CIDEr and SPICE) of a model trained on a distilled version of each of the scores in isolation, compared to the combined *ICC* metric which outperforms both.

***ICC* Model Component Ablations.** In Table 6, we ablate the effect of various design choices in the *ICC* pipeline by evaluating their effects on caption concreteness prediction (using the benchmark described in Section 3.2). In particular, we test different LLM sizes (Zhang et al., 2024; Geng and Liu, 2023) in the SBA pipeline, different captioning model architectures in the VBA pipeline, and the similarity measure used in each pipeline (edit distance vs. BERTScore). To identify the effect of each component, we evaluate SBA and VBA predictions in isolation (without combining or distilling them). As is seen in the table, our chosen LLM and captioning model perform comparably

		COCO			Flickr		
Filt.	Size	R@1	R@5	R@10	R@1	R@5	R@10
Rand.	100k	5	15.4	23.3	10.6	31.5	42.6
CLIP	100k	2.1	7.5	12.4	5.7	17.1	26
CA	100k	5.2	15.8	24.1	11.3	32.2	43.8
TMS	100k	6.5	19.5	28.8	14.9	37.1	49.5
PAC	100k	4.8	14	21.2	9.2	24.7	35.5
<i>ICC</i>	100k	14.4	34.5	45.7	32.6	62.7	73.5
Rand.	200k	9.6	25.5	36.2	21.1	48.9	61.8
CLIP	200k	6.9	10	15.8	6.9	20.9	30.9
CA	200k	8.8	24.4	35.1	20.8	48.6	61.2
TMS	200k	8.2	23	32.8	17.8	43.4	56.3
PAC	200k	6.5	17.7	26.3	12.9	31.1	42.9
<i>ICC</i>	200k	15.5	35.8	47.6	33.6	63.2	74.5
Rand.	500k	8	22.2	32.5	17.4	42.1	55.4
CLIP	500k	5.3	16	23.9	11.5	30	42.7
CA	500k	8.2	22.6	32.4	17	43.3	56.7
TMS	500k	10	26.3	37.2	20.3	46.8	60.5
PAC	500k	8.8	23.5	33.9	17.8	40.4	53
<i>ICC</i>	500k	14.6	34.9	47	30.6	60.9	72.9

Table 2: **Representation learning results over different filtered dataset sizes.** We perform text-to-image retrieval evaluation over MS-COCO and Flickr30K for different filtering schemes of LAION-400M with varying dataset sizes. We compare our performance (*ICC*) to various filtering baselines: Rand. indicates selecting random samples from LAION-400M, CLIP indicates filtering by top CLIPScore, CA indicates Complexity and Action filtering, TMS indicates filtering with T-MARS and PAC indicates filtering with PACScore. Best results are in **bold**.

to the alternative models tested, showcasing the robustness of the VBA and SBA across model sizes. Moreover, while the simple edit distance similarity measure performs acceptably for the SBA pipeline, the BERTScore similarity measure produces significantly better correlations in the VBA pipeline, matching the intuition that the VBA is inherently lossy with respect to the precise form of texts and must rely on a more semantic measure to properly detect abstract sentences.

5 Related Work

Evaluating Text Concreteness. Word concreteness is a topic of interest in cognitive science (Paivio et al., 1968; Richardson, 1975; Schwanenflugel, 2013; Khanna and Cortese, 2021), and a number of works have studied auto-

Method	Word Conc.			Caption Conc.		
	ρ	ρ_s	τ	ρ	ρ_s	τ
CLIP-SP	0.6	0.62	0.44	0.34	0.33	0.25
aveCLIP	0.55	0.56	0.39	0.29	0.28	0.22
GPT-3.5	0.55	0.56	0.44	0.44	0.48	0.4
GPT-4o	0.78	0.79	0.64	<u>0.57</u>	<u>0.57</u>	<u>0.49</u>
<i>ICC</i>	<u>0.75</u>	<u>0.75</u>	<u>0.55</u>	0.73	0.75	0.6

Table 3: **Concreteness evaluation on single-word and caption-level texts.** Correlation (in absolute value) is measured using Pearson ρ , Spearman ρ_s , and Kendall τ coefficients. Best result are in **bold**, second best are underlined.

	ρ	ρ_s	τ
Before Distillation	0.65	0.6	0.46
After Distillation	0.72	0.75	0.6

Table 4: **Distillation Effect on Caption Concreteness Correlation.** We show correlations to ground-truth annotated caption concreteness scores before and after distillation. The ‘‘After Distillation’’ row corresponds to our final *ICC* score.

matic 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 caption or string level. Shi et al (2019) define concreteness of constituents by matching them to images for learning syntactic representations without explicit supervision; as was later shown, the signal of noun concreteness plays a key role in the model’s syntactic predictions (Kojima et al., 2020). Most similar to us is Wu and Smith (2023), who generate multiple images for each caption and average the CLIP similarity scores over all the images to produce a caption-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

Method	COCO		NoCaps	
	CIDEr	SPICE	CIDEr	SPICE
SBA	17.8	5.9	15.1	3.3
VBA	29.8	9.4	27.8	5.8
ICC	30.9	9.7	28.2	6

Table 5: **Score Ablations** We ablate the importance of using scores obtained from both the SBA and VBA pipelines over 200k samples dataset that was filtered using the different scores.

pairs (Schuhmann et al., 2022), de-duplicating semantically similar content (Abbas et al., 2023), and removing samples with text that overlap with the image (Maini 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; Lai et al., 2023; Nguyen et al., 2023). By contrast, we do not require modifying the given dataset and identify 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.

6 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 in an unsupervised manner and find that this highly correlates with human concreteness judgments. Our results demonstrate that *ICC* is effective at selecting a core of high-quality image-caption samples from web-scale multimodal datasets for training models in the resource-constrained setting. We foresee the use of *ICC* in additional tasks requiring the curation of web-scale multimodal data, where high-quality, visually-concrete text is needed.

Limitations

While our method manages to detect visually concrete captions well, it lacks sensitivity to grammatical structure, which might cause it to label oddly phrased captions as concrete. For instance, consider the caption: “a computer near a tree with a boy next to a table with a keyboard”. This cap-

Caption Concreteness					
Pipe	Model Part	Sim.	ρ	ρ_s	τ
SBA	TinyLLaMa-1.1B	ED	0.59	0.58	0.45
SBA	OpenLLaMa-3B	ED	0.57	0.56	0.43
* SBA	LLaMa-2-7B	ED	0.53	0.51	0.48
SBA	TinyLLaMa-1.1B	BSc	0.57	0.56	0.43
SBA	OpenLLaMa-3B	BSc	0.56	0.55	0.42
SBA	LLaMa-2-7B	BSc	0.57	0.56	0.43
VBA	BLIP-Base	ED	0.43	0.4	0.31
VBA	BLIP-Large	ED	0.43	0.36	0.27
VBA	BLIP-2	ED	0.44	0.41	0.31
VBA	BLIP-Base	BSc	0.6	0.6	0.46
VBA	BLIP-Large	BSc	0.58	0.56	0.43
* VBA	BLIP-2	BSc	0.6	0.58	0.45

Table 6: **Ablations over VBA and SBA Design Choices.** We ablate the effect of the LLM used in the SBA pipeline and the captioning model used in the VBA pipeline, as well as the text similarity measure, on the correlation to the ground-truth concreteness annotations. Note that here we measure correlation to each model of the pipelines (VBA and SBA) used in isolation. BSc and ED refer to BERTScore and edit distance respectively. We report the Pearson ρ , Spearman ρ_s , and Kendall τ correlation coefficients. Our default settings are indicated with a prepended *.

tion is highly concrete and gets a high *ICC* score of 1.0. However, removing all object relations from the caption produces the following: “computer tree boy table keyboard” which results in a relatively minor decrease of the *ICC* score to 0.89. Such low-quality captions might have a negative impact on tasks such as image captioning where the model must learn to output grammatically correct English sentences which should ideally describe relevant fine-grained relations between entities. We hypothesize that this behavior stems from the dataset used to train the distillation model (CC3M) which is not likely to include such oddly phrased captions, and so these non-grammatical structures are not learned. We hypothesize that training over a dataset with higher caption diversity will likely alleviate this issue.

In addition, due to limited computational resources, our experiments were conducted on a relatively small scale of 8 million sample initial training dataset based on LAION-400M. We expect that increasing the scale and the filtered dataset proportionally will result in a performance improvement in the downstream model performance.

However, we leave verifying this as well as testing the effect of ICC filtering on other downstream tasks such as VQA and caption ranking to future work.

Finally, while our method detects and filters an important category of noise in multimodal datasets, we note that abstract captions such as those in Figure 2 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.

Acknowledgements

This work was partially supported by the KLA Foundation and Google. We thank Yonatan Bitton and Keren Ganon for their helpful feedback.

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*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. No-caps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.
- 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.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- 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 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*.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.
- 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.
- Xinyang Geng and Hao Liu. 2023. [Openllama: An open reproduction of llama](#).

- 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*.
- Maya M Khanna and Michael J Cortese. 2021. How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. *Memory*, 29(5):622–636.
- Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. 2020. What is learned in visually grounded neural syntax acquisition. *arXiv preprint arXiv:2005.01678*.
- Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jilong Shan, Chen-Nee Chuah, Yinfei Yang, et al. 2023. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*.
- 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.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- 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*.
- Pratyush Maini, Sachin Goyal, Zachary C Lipton, J Zico Kolter, and Aditi Raghunathan. 2023. T-mars: Improving visual representations by circumventing text feature learning. *arXiv preprint arXiv:2307.03132*.
- Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*.
- Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- 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.
- John TE Richardson. 1975. Concreteness and imageability. *The Quarterly Journal of Experimental Psychology*, 27(2):235–249.
- Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6914–6924.
- 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.
- Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. *arXiv preprint arXiv:1906.02890*.
- 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*.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. *arXiv preprint arXiv:2402.18091*.
- Si Wu and David Smith. 2023. [Composition and deformation: 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.
- Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. [Tynyllama: An open-source small language model](#).
- 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 Standardizing 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 for both the SBA and VBA. In Figure 5, we show the distribution of the edit-distance based reconstruction similarities of the SBA before and after standardization per caption length. We can see in Figure 5a 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 5b 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.2 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 (Liu 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 6. We regress over these samples to get the optimal weights.

A.3 Caption Concreteness Benchmark Distribution

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 curate our captions to achieve a balanced distribution of concreteness scores, as illustrated in Figure 7. As seen there, the concreteness of the benchmark’s captions is evenly distributed between abstract and concrete concepts.

A.4 Zero-Shot CLIP Concreteness Score

We adapt the Stroop Probing method (Alper et al., 2023) for estimating text concreteness. While

Alper et al. (2023) test this on single words, we adapt this method to captions by replacing the empty slot in prompts with a caption rather than a single word. We use their prompts, omitting those which do not match the context of an entire caption being inserted in the masked slot (i.e., omitting the prompts “Alice giving the [*] to Bob” and “Bob giving the [*] to Alice”).

A.5 GPT Prompts

The following prompts were used to extract concreteness scores for image captions⁴ from GPT-3.5 and GPT-4o:

System: “You are an expert visual reasoner, capable of understanding the visual concreteness of image captions. A visually concrete caption is a caption that is highly visual, and can be vividly imagined.”

User: “Provide a numerical score on a scale of 1-5, when 1 is non-visual and 10 is highly visual caption for the following caption : <caption>. Only provide the numerical score and nothing else.”

Note that we experimented with three different ranges of [1-N] of concreteness scores in our prompts: N=3, N=5 and N=10. We found that N=5 yielded the best results.

A.6 aveCLIP Word Concreteness

Since aveCLIP requires generating many images per word or caption, we found that running aveCLIP over the entire word concreteness dataset is not feasible due to runtime constraints. Therefore, we evaluate its performance on a random subset of 150 words/captions.

A.7 Training Hyperparameters and Additional Information

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

VBA Text-To-Image. For the image generation of the diffusion model in the VBA, we use guidance scale of 9 and 20 inference steps.

⁴To get the concreteness scores of words, we used the same prompts with “word” instead of “caption” in the appropriate places.

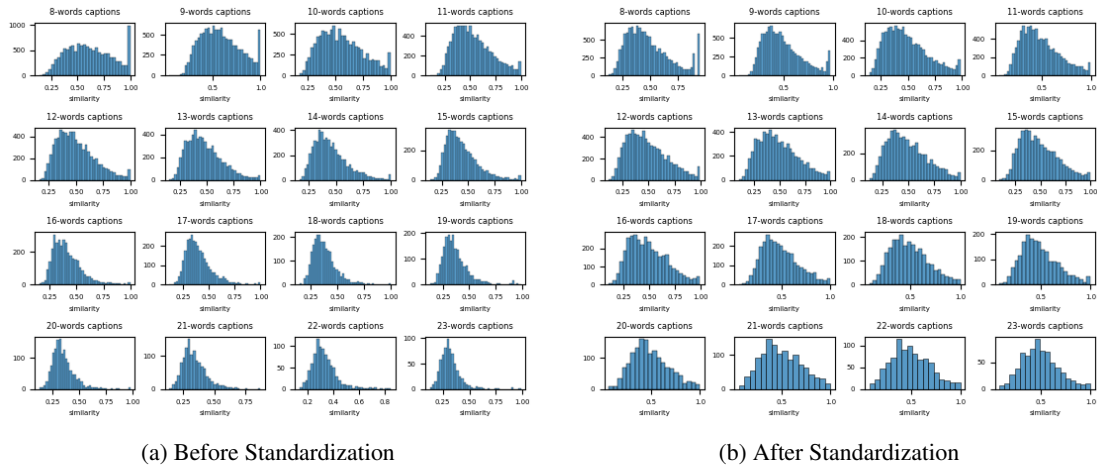


Figure 5: **Standardizing by caption length.** We show the reconstruction similarity scores of SBA for each caption length before standardization (in 5a) and after standardization (in 5b).

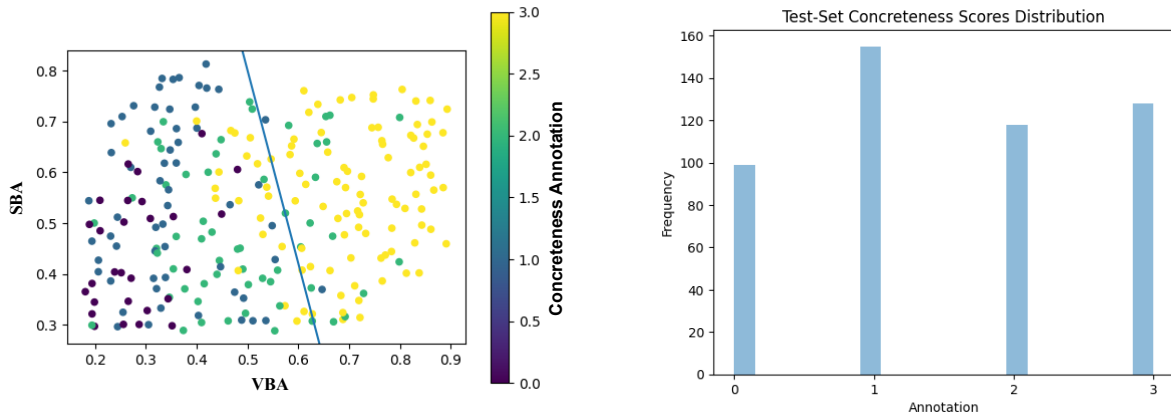


Figure 6: **Finding the Optimal Weights.** We measure the optimal combination of the two scores with respect to ground-truth concreteness annotations.

Figure 7: Distribution of annotated concreteness scores in our manually labeled test set of 500 captions. All samples are from LAION-400M. Annotations range from highly abstract (0) to highly concrete (3).

A.8 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 Hub⁵. 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:

- distilroberta-base

⁵<https://www.huggingface.co/models>

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

A.9 Finding the Score Combination Parameters

To compute the combination parameters of the SBA and VBA scores, we label 244 captions, sampled uniformly over VBA and SBA scores, with concreteness scores in the range 0–3. We use logistic regression to find the parameters a, b, c of $\sigma(a \cdot VBA + b \cdot SBA + c)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$

is the sigmoid function, such that the output will approach 1 for concrete captions and 0 for abstract ones. We label concrete captions as captions with concreteness above the median score in the labeled dataset and abstract captions as captions with a score below this median. We visualize the annotated samples and the regression line $a \cdot VBA + b \cdot SBA + c = 0$ in Figure 6. The parameters found and used in our *ICC* are $a = 13.2$, $b = 3.6$ and $c = -9.4$. As seen in the figure, both scores contribute to the optimal predicted concreteness score, validating the importance of using both SBA and VBA components together in our full pipeline.

Appendix B Additional Qualitative Examples

We visually show examples of each of the scores’ weaknesses and the way they compliment each other. In Figure 8, 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. These cases mainly stem from 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 manner, we show in Figure 9 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 compensates 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 Figures 8–9, it can be observed that *ICC* reflects the advantages of both pipelines by generating low scores for abstract captions, and high scores for concrete ones in a consistent manner.




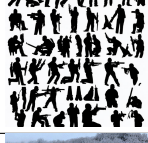
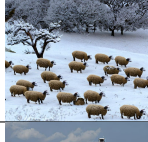



Input caption	SBA reconstructed caption	VBA re-constructed caption	VBA bottleneck image	SBA	VBA	ICC
a nurse mopping a surgeon's brow during an operation in an operation pub	a nurse wiping the brow of a surgeon during an operation in an operating room	two people in protective gear		0.77	0.25	0.72
bougainvillea 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 hill-top	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 8: **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 unsuccessful 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 9: **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.