Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved With Text

Wanrong Zhu[♣*] **Jack Hessel**[♡*]
Anas Awadalla[♠] **Samir Yitzhak Gadre**[◇] **Jesse Dodge**[♡] **Alex Fang**[♠]
Youngjae Yu[†] **Ludwig Schmidt**[♠♡‡] **William Yang Wang**[♣] **Yejin Choi**[♠♡]
[♣]University of California, Santa Barbara [♡]Allen Institute for Artificial Intelligence
[♠]Paul G. Allen School of Computer Science, University of Washington
[◇]Columbia University [†]Yonsei University [‡]LAION
`https://github.com/allenai/mmc4`

Abstract

In-context vision and language models like Flamingo [2] support arbitrarily interleaved sequences of images and text as input. This format not only enables few-shot learning via interleaving independent supervised (image, text) examples, but also, more complex prompts involving interaction between images, e.g., "What do image A and image B have in common?" To support this interface, pretraining occurs over web corpora that similarly contain interleaved images+text. To date, however, large-scale data of this form have not been publicly available.

We release Multimodal C4 (`mmc4`), an augmentation of the popular text-only `c4` corpus[2] with images interleaved. We use a linear assignment algorithm to place images into longer bodies of text using CLIP features [20], a process that we show outperforms alternatives. `mmc4` spans everyday topics like cooking, travel, technology, etc. A manual inspection of a random sample of documents shows that a vast majority (90%) of images are topically relevant, and that linear assignment frequently selects individual sentences specifically well-aligned with each image (78%). After filtering NSFW images, ads, etc., the corpus contains 103M documents containing 585M images interleaved with 43B English tokens.

1 Introduction

In-context learning [7] enables sequence models to adapt to new tasks without any parameter updates. By interleaving a few supervised examples in a prompt, few-shot learning can be formatted as a next-token prediction task, i.e., $x_1, y_1, x_2, y_2, \ldots, x_n$ is input to predict $\hat{y}_n$. Some image+text models also support in-context learning via interleaving of images/text jointly. Prior experiments [2] suggest that performant multimodal in-context learning is dependent upon pretraining on similarly interleaved sequences of images and text (rather than single image/caption pairs). However, such a large-scale corpus has not been made publicly available.

To address this, we introduce Multimodal C4 (`mmc4`), a public, billion-scale image-text dataset consisting of interleaved image/text sequences. `mmc4` is constructed from public webpages contained in the cleaned English `c4` corpus. In addition to standard preprocessing steps like deduplication, NSFW removal, etc., we place images into sequences of sentences by treating each document as an instance of a bipartite linear assignment problem, with images being assigned to sentences (under the

[*] equal contribution; work partly conducted while Wanrong Zhu was an intern at AI2.
[2] `https://www.tensorflow.org/datasets/catalog/c4`

	# images	# docs	# tokens	Public?
M3W (Flamingo) [2]	185M	43M	-	×
Interleaved training data for CM3 [1]	25M	61M	223B	×
Interleaved training data for KOSMOS-1 [13]	$\leqslant$ 355M	71M	-	×
Multimodal C4 (mmc4)	585M	103M	43B	✓
Multimodal C4 fewer-faces (mmc4-ff)	385M	79M	34B	✓
mmc4 core (mmc4-core)	30.5M	7.4M	2.5B	✓
mmc4 core fewer-faces (mmc4-core-ff)	22.9M	5.6M	1.8B	✓

Table 1: Comparison of mmc4 with other interleaved image/text pretraining corpora. In addition to the full version of the dataset, we also release 1) fewer-faces subsets, which aim to remove all depicted human faces; and 2) "core" subsets, result from more stringent filtering.

constraint that each sentence is assigned at most one image). We show that applying CLIP ViT-L/14 [20] to estimate bipartite weights in a zero-shot fashion results in state-of-the-art performance on intra-document alignment benchmarks, and then apply this process to 100M+ documents to construct mmc4.

We explore mmc4, showing that: 1) the text and images in the corpus span expected everyday topics like cooking and travel; 2) filters like NSFW/ad removal work with high accuracy; and 3) the resulting images are relevant to the associated documents, and often, appropriately aligned to the most-relevant individual sentence. We conclude by discussing initial use-cases of mmc4, including OpenFlamingo [3],[3] an open source version of Flamingo [2]. Initial ablations show that training on the sequences of mmc4 enables few-shot, in-context adaptation to image captioning datasets.

2 Related dataset work

Most million/billion-scale, public multimodal pretraining datasets consist of images paired with their literal descriptions, e.g., LAION-2B [22], CC-12M [8], YFCC100M [28]. However, literal description is only one of many ways images can relate to text on the web [17]. mmc4 aims to capture a broader range of these relationship types. Some web datasets situate images in longer bodies of text, e.g., The Wikipedia-based Image Text Dataset [26] (11.5M images), but do not directly cover multi-image/multi-sentence interleaving. Table 1 provides summary statistics of other large-scale interleaved pretraining datasets. mmc4 contains more images than prior non-public datasets. [5] highlight risks associated with web-scale multimodal data. In addition to curation steps detailed in § 3 and the release considerations in § 3.1, we're hopeful that mmc4's availability can enable more open auditing+critique of interleaved corpora compared to previous private training sets. Models trained on mmc4 inherit its risks; we selected the widely-adopted c4 corpus as a starting point in part because there are existing auditing efforts on the text-only corpus, see § 3; and [19] for more discussion of transparency.

3 Data Curation Process

Initial data collection. Multimodal C4 is an expansion of the text-only c4 dataset [21], which was created by taking the April 2019 snapshot from Common Crawl[4] and applying several filters with the intention of retaining high-quality, natural English text. Each document in c4 consists of the text scraped from one URL. The full c4 dataset has 365M documents and 156B tokens, covering many domains [11]; it was first used to train T5 [21]. We built the mmc4 dataset on top of c4 because: 1) c4 is a web-scale dataset widely adopted as a pre-training corpus [21, 25, 9, 29, 27]; 2) c4 is constructed from web pages, which frequently contain multimedia content like images: a multimodal sequence version is a natural extension; and 3) c4-en,[5] the specific underlying subset from which

[3]https://github.com/mlfoundations/open_flamingo

[4]https://commoncrawl.org/

[5]https://www.tensorflow.org/datasets/catalog/c4#c4en_default_config

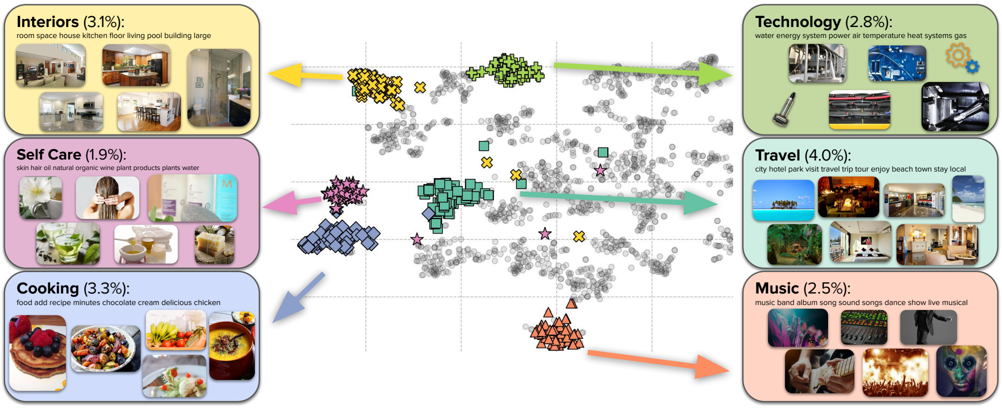


Figure 1: A T-SNE [30] projection of LDA [6] topic clusters from a random sample of 22K documents from `mmc4`; `mmc4` spans a variety of everyday topics, e.g., cooking, technology travel, etc. For 6 selected topics, we also show a sample of most-central images to the topic according to CLIP ViT-L/14 [20].

we construct `mmc4` has already been processed with several data-cleaning steps (including English-language identification by `langdetect`[6] with at least 0.99 confidence; text deduplication removing duplicate three-sentence spans + placeholder text like "lorem ipsum"; and removal of any document containing any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words").[7] See [21] for more information about the text-only `c4`. Importantly, by building on the popular text-only `c4`, prior text-only documentation efforts [11] can provide insight about potential biases and risks that could arise when training on our multimodal extension. We use the NLTK [4] sentence tokenizer to chunk each `c4` document into a list of sentences.

Gathering images. We first retrieve the original webpages for each document in the `c4-en` dataset from the Common Crawl version `2019-18`, which is the default version for `c4`. Next, we extract the URLs for downloadable images from the raw WAT files. We restrict the image extension to either `png/jpeg/jpg`, and exclude image URLs that contain the following tokens: {`logo, button, icon, plugin, widget`}. We attempt to download from these URLs, and resize images to a maximum dimension of 800px. We eliminate any `c4` documents that do not contain valid, downloadable images at the time of collection (mid-to-late 2022). The starting point after this step is 115M documents and 1.37B images.

De-duplication+small resolution. We next run duplicate image detection using opennota's `findimagedupes`[8] which uses `phash` to identify visually similar images.[9] We keep only one copy of an image if multiple versions are detected within the same document. We also remove images with more than 10 duplicates in a sample of 60K images. We discard images with a width or height smaller than 150px; this accounts for many small icons, e.g., navigation buttons. We discard images with an aspect ratio of greater than 2 or less than 0.5; this accounts for many banner-like ads. In a manual sample of 3.7K images that survive this (and the NSFW) filter, 91 images (2.5%) were identified as ads potentially unrelated to document contents.[10]

[6] `https://pypi.org/project/langdetect/`

[7] `https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words`

[8] `https://gitlab.com/opennota/findimagedupes`

[9] We use a more aggressive de-duplication threshold of 5 compared to the default library setting of 0; this removes roughly 10M additional images. While some duplicates survive this process, we qualitatively found a threshold of 5 to be an appropriate balance of false positives/negatives.

[10] The delineation between an "irrelevant advertisement" and a "relevant image" is inexact: for example, we discovered images advertising specific, small events, e.g., ones hosted by a fishing club within a city (this type of image was not included in this count). We later assess advertisement-ess in the context of the text of documents, rather than assessing based on the image alone.

	MSCOCO		Story-DII		Story-SIS		DII-Stress		RQA		DIY	
	AUC	p@1	AUC	p@1	AUC	p@1	AUC	p@1	AUC	p@1	AUC	p@1
Random	49.7	5.0	49.4	19.5	50.0	19.4	50.0	2.0	49.4	17.8	49.8	6.3
Hessel et al. (2019) [12]	98.7	91.0	82.6	70.5	68.5	50.5	95.3	65.5	69.3	47.3	61.8	22.5
Li et al. (2021) [16]	99.3	**97.6**	85.5	77.2	70.2	53.1	–	–	–	–	–	–
CLIP ViT-L/14 (Zero Shot)	**99.4**	95.7	**92.8**	**93.9**	**79.1**	**73.3**	**98.7**	**93.0**	**80.7**	**70.7**	**74.0**	**57.6**

Table 2: Performance on single document image-text benchmarks from [12] (higher=better in all cases). Applying CLIP ViT-L/14 in a zero-shot fashion [20] produces better within-document alignments compared to prior methods which rely on fine-tuning.

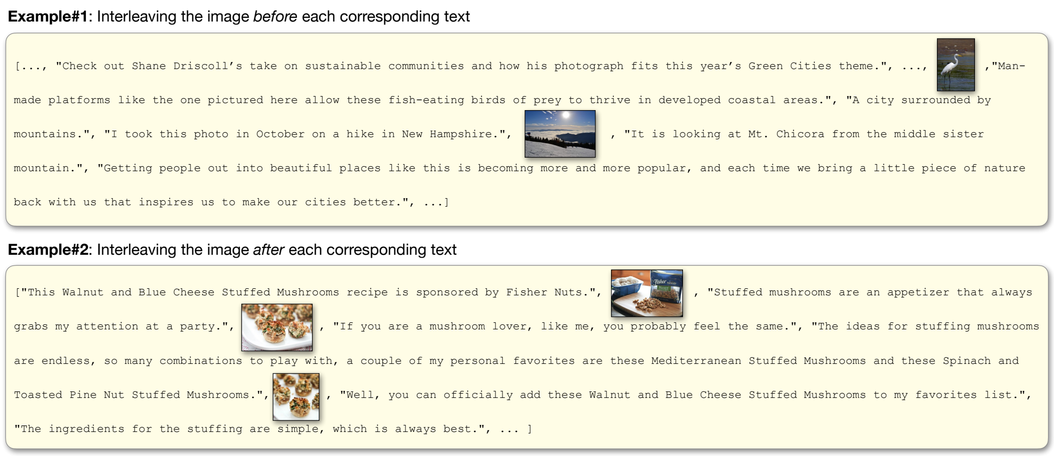


Figure 2: Two example image+text documents from mmc4. Following Flamingo [2], during training, images can be interleaved before or after their assigned sentences. More example documents are given in Appendix C.2.

Discarding NSFW images. We run an NSFW binary image classifier on each image, which is trained on the dataset introduced in LAION-2B [22]. The model is a 4-layer MLP trained over image features extracted from OpenAI's CLIP ViT-L/14 [20] and achieves 97.4% accuracy on the NSFW test set. We discard cases with a model-predicted NSFW probability over 0.1, which removes approximately 10% of remaining images. In a manual sample of 3.7K images that survive this filter in mmc4, we discovered zero NSFW images.

Aligning images and sentences. After collecting a set of images for each document, we now describe our intra-document alignment process to interleave the collected images with the sentences. Given that the scope of the images and sentences may be different – the image set is collected from the whole webpage, while the sentence list is subject to preprocessing within the c4 dataset and thus may not represent the complete content of the webpage – we did not rely on Document Object Model placements in the raw HTML to establish the alignment between images and sentences in each document. Instead, to associate each image with a sentence, we consider each document as an instance of a bipartite assignment problem [15, 12], and use CLIP ViT-L/14 compute pairwise similarities between all sentences/images on a single page. Then, we discard images without at least a 0.15 CLIP cosine similarity to at least one sentence in the document. Finally, we use [14] to compute a bipartite assignment of images to sentences, under the constraint that each sentence can only be assigned a single image.[11] Table 2 shows that this zero-shot application of CLIP ViT-L/14 for within-document matching surpasses prior competitive, fine-tuned methods on image-text alignment benchmarks from [12] (we also distribute the raw intra-document similarity matrices with mmc4 so alternate assignment methods can be explored). Figure 2 illustrates two example documents with the images interleaved before or after the assigned sentences.

[11]For documents with more images than sentences, after assigning an image to each sentence, we assign according to max similarity.

3.1 Considerations for data release

mmc4 contains all images that survive the previously described filters. In addition to the full version of the corpus, we construct two additional types of subsets.

3.1.1 Fewer Faces (mmc4-ff)

Like the text-only version of c4, mmc4 may contain webpages with personal information that individuals had not explicitly intended to make available for model training. For an initial public release, we make a version of mmc4 available, mmc4-ff (ff stands for "fewer faces") that aims to remove images containing detected faces.

Removing images with detected faces. To detect faces at billion-scale with the intent of removing them from the dataset, we first run RetinaFace[10][12] over a sample of 60K images with the default settings. This detector runs at a high resolution and would be computationally prohibitive to run in full precision for the whole corpus; it produces detailed localization information about the coordinates of each face in each image (which we discard). Using an 80/20 train/test split, we train a cross-validated logistic regression over CLIP ViT-L/14 features to predict whether or not RetinaFace detects a face: this classifier is several orders of magnitude faster compared to RetinaFace. This approximation performs well: we choose a confidence cutoff that achieves 95% recall[13] for the label "RetinaFace detected any face" over the test set while preserving 65% of the original images.

Manual sample-based face image risk assessment. We performed a manual verification of face removal. In a random sample of 912 images that pass all filters including the "no faces" filter, 23 (2.5%) images arguably contain a mostly-un-obscured human face. In most cases (12/23), faces are very low resolution, e.g., a 150x150px image of a crowd of people from a distance, where each face accounts for 3x4 pixels, or are motion shots where the face is blurred. In one case, the face is Marilyn Monroe's as depicted in art on a wall. In 6 cases, there is a plausibly identifiable face depicted: in 2 cases, these are models posing in ads; in 1 case, there is a low resolution image of politicians giving a speech; in 2 cases, the faces are obscured; in 1 case, a passerby was caught in the background of a city photograph and could feasibly be individually identified. Overall: the rate of unobscured, high-resolution, identifiable faces in mmc4-ff is low.

3.1.2 Core (mmc4-core)

Early conversations with some model developers revealed a desire to work with a smaller subset of the corpus as an initial step. We thus additionally release core versions of mmc4 (and mmc4-ff), which apply even more stringent filtration criteria. The aim of core is to identify a "higher-precision" subset of documents that: 1) have a minimum/maximum number of sentences/images per document; 2) pass an even stricter deduplication step; and 3) have a higher image-text similarity. Hyperparameters[14] are selected heuristically and are balanced to downsize the original corpus by an order of magnitude.

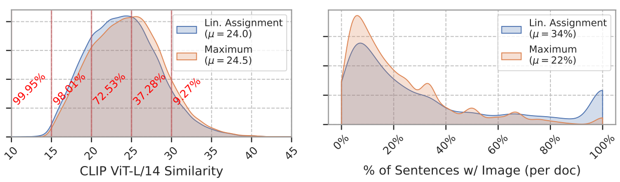
4 Exploring mmc4

Statistics. Table 1 gives basic summary statistics of mmc4 (and fewer-faces/core subsets) compared to some other interleaved image/text corpora. Overall, the full version of mmc4 is larger than prior non-public datasets across axes like number of images/number of documents. In addition, the various subsets of the corpus offer trade-offs between privacy, image/text similarity thresholds, etc. Figure 5 gives details about the mean/median number of images/sentences in each document (mean/median # sent.=2.0/5.7; # im = 13.0/24.3) based on a random sample of 22K documents.

[12] As implemented by [23, 24] available from https://github.com/serengil/retinaface.

[13] RetinaFace is not perfectly accurate, so selecting a more aggressive threshold (e.g., 99.99%) would not necessarily result in significantly fewer face-containing images removed.

[14] Min/max number of sentences: 4/40; min/max number of images 2/15; findimagedupes applied with a threshold of 10; documents are required to have at least 75% of image assignments have CLIP ViT-L/14 similarity of greater than 25.



(a) CLIP sim is similar between lin. assignment + max. In red: percent of images remaining at various CLIP thresholds.

(b) Lin. assignment results in a higher percentage of sentences being associated with an image.

Figure 4: Using linear assignment results in comparable image-text similarities to max assignment, but the former spreads images much more evenly, e.g., the per-document mean percent of sentences with an associated image increases from 22% to 34%.

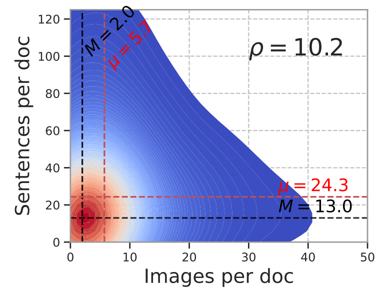


Figure 5: Distribution of images and sentences per document; the median document has 2 images/13 sentences. Documents with more sentences tend to have more images, but the correlation is weak (Spearman $\rho = 10.2$).

Image-text similarity. Figure 4 provides detail about the linear assignment process compared to a "max" assignment alternative, where each image is simply assigned to its maximally CLIP-similar sentence. The linear assignment process slightly decreases the average CLIP similarity between images/sentences (from 24.5 → 24.0), but significantly more evenly "spreads" images throughout the documents: per-document, the mean percentage of sentences with an associated image rises from 22% → 34%.

Topic-based assessment. We ran LDA [6] as implemented by Mallet [18] on a random sample of 22K documents from `mmc4` with $k = 30$ topics. The resulting clusters span a broad set of topics like cooking, communities, travel, music, art, etc. Figure 1 shows some example LDA topic clusters.[15] In addition, we explore a sample of the images most associated with the corresponding topic,[16] finding that, in general, image topic clusters align with qualitative expectations.

Manual verification of image relevance+properties. We randomly sample 200 documents from `mmc4` with the goal of assessing how relevant the images contained in the document are to the assigned sentences and to the document as a whole. Table 3 shows the results on the 799 images contained in the 200 documents. 89.5% of all examined images are topically related to the corresponding document, and 77.6% images are well-aligned to the assigned sentences within each document.[17] We also assessed several other factors, finding that: 1) 22.3% contain recognizable human faces; 2) 3.3% contain recognizable watermarks; 3) 4.1% are related to logos;[18] 4) 2.1% are related to advertisements; and 5) 1.4% are duplicated with other images in the same document. More discussion of images with watermarks, ads/logos, etc. can be found in Appendix C.1.

5 OpenFlamingo: An Early Application of `mmc4`

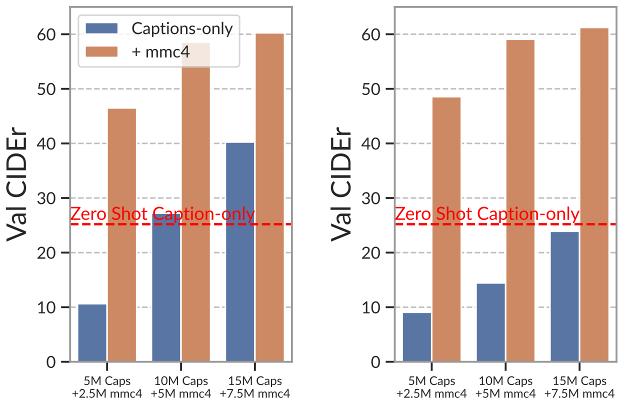
The first publicly available model to be trained on `mmc4` is OpenFlamingo [3]. We run ablations on a small version of OpenFlamingo (3B: backbone = OPT-1.3B [32] language model and CLIP ViT-L/14 [20] vision model) to compare direct training on image captions (LAION-2B [22]) to the interleaved

[15] A full list of topics and their frequencies according to the model is in Appendix A

[16] We compute the mean CLIP ViT-L/14 image vector for each topic by associating each image in a document the document's most common topic; then, we compute the mean image vector per topic. Finally, cosine similarity to this mean vector is used to identify the "most topically central" images per-topic.

[17] The alignment between an image and its assigned sentence is a qualitative criterion. We consider an image-sentence pair to be "well-aligned" when the visual elements of the image have a direct and relevant relationship with the text. This can include instances where the image depicts the context or content of the sentence, or where there is a plausible literal overlap between the text and the image, etc.

[18] The logos can be website logos, commercial logos used by businesses or companies to represent their brand or product, or logos for organizations or events. In all cases, the label is assigned if the logo is the primary focus of the image.



(a) 4-shot

(b) 8-shot

Figure 6: Few shot, in-context MSCOCO captioning performance of OpenFlamingo-3B when training on just captions from LAION-2B vs. mixing in mmc4-core sequences. The model trained on mmc4 sequences is able to generalize to MSCOCO-style captions more effectively vs. the model trained just on LAION-2B image/caption pairs. (Zero shot caption-only=15M caption LAION-2B model)

	% of 799 images
Topically-related	89.5%
Sentence-aligned	77.6%
Has face?	22.3%
Has watermark?	3.3%
Logo-related	4.1%
Ads-related	2.1%
Duplicated	1.4%

Table 3: Results of manual verification of 200 randomly sampled documents containing 799 images. A majority of images are topically relevant and well sentence-aligned. The rate of watermarks, ads, duplicates, etc. is low.

sequences of `mmc4-core`. To flatten `mmc4` documents to training sequences,[19] We: 1) sample a 256 token sub-sequence from each training document; 2) discard images with CLIP image-text similarity less than 20; 3) discard sequences that contain no images after filtering; 4) discard images if there are more than 5 in the resulting sequence.[20] As in [13] we randomly drop sequences with a single image to increase multi-image sequences we sample.

Validation CIDEr [31] results for COCO image captioning are in Figure 6. For 4/8-shot in-context learning settings, the model trained on `mmc4-core` shows 20-30 CIDEr point improvements. The performance of OpenFlamingo-3B trained on just 5M captions/2.5M `mmc4` sequences also exceeds a zero-shot application of OpenFlamingo-3B trained on much more data (15M LAION-2B captions); this provides additional evidence that the interleaving in-context setup enables adaptation to MSCOCO-style captions. The performance of the captions-only OpenFlamingo-3B model degrades from 4-shot to 8-shot learning presumably because these longer sequences are significantly different from the single image/captions it's seen at training time.

6 Conclusion

We introduce `mmc4`, a corpus of 585M images interleaved in 43B English tokens from the popular `c4` dataset. Models trained on image/text sequences from `mmc4` can more effectively perform multimodal in-context learning compared to models trained on single image/captions. We expect interleaving will be important not only for few-shot learning, but also for more diverse multimodal language technologies wherein users may seek to converse with agents with and about visual content in new ways. Future work includes:

1. More precise empirical evaluation of in-context abilities: can models really reason across images/texts in a prompt in flexible ways, or are they limited to interleaved and independent supervised examples?
2. Data scaling: is the performance of in-context vision+language learning bottlenecked by the availability of large-scale interleaved corpora? Or is improved single-modal pretraining sufficient to un-bottleneck multimodal models?

[19]Future work would be well-suited to investigate the impact of various flattening schemes on downstream performance; the method described here is just one possible method.

[20]Similar to [2], we find that training on a maximum of five image sequences can be sufficient for OpenFlamingo models to generalize to 32 shots during inference.

3. Instruction tuning: while interleaving of independent supervised image+text examples enables in-context learning, training an instruction-following multimodal model directly for this case is a promising alternative.

7 Acknowledgements

We thank the OpenFlamingo team, Sangho Lee, and Jiasen Lu for the helpful discussions, and for being early adopters of `mmc4`. In addition, we thank Jingkang Yang for helpful discussions inspiring `mmc4-core`. We thank Stability AI for the compute for the OpenFlamingo experiments. This work was supported in part by DARPA MCS program through NIWC Pacific (N66001-19-2-4031), the NSF AI Institute for Foundations of Machine Learning (IFML, CCF-2019844), Open Philanthropy, Google, and the Allen Institute for AI.

References

[1] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. Cm3: A causal masked multimodal model of the internet. *ArXiv*, abs/2201.07520, 2022.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[3] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[11] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Jack Hessel, Lillian Lee, and David Mimno. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *EMNLP*, 2019.

[13] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045, 2023.

[14] Roy Jonker and Ton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. In *DGOR/NSOR: Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR*, pages 622–622. Springer, 1988.

[15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[16] Zejun Li, Zhongyu Wei, Zhihao Fan, Haijun Shan, and Xuanjing Huang. An unsupervised sampling approach for image-sentence matching using document-level structural information. In *AAAI*, 2021.

[17] Emily E Marsh and Marilyn Domas White. A taxonomy of relationships between images and text. *Journal of documentation*, 59(6):647–672, 2003.

[18] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002.

[19] Alex Mei, Michael Saxon, Shiyu Chang, Zachary C Lipton, and William Yang Wang. Users are the north star for ai transparency. *arXiv preprint arXiv:2303.05500*, 2023.

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

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[23] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.

[24] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.

[25] David So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Searching for efficient transformers for language modeling. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[26] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR*, 2021.

[27] Jun Suzuki, Heiga Zen, and Hideto Kazawa. Extracting representative subset from extensive text data for training pre-trained language models. *Information Processing & Management*, 60(3):103249, 2023.

[28] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[29] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Huai hsin Chi, and Quoc Le. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239, 2022.

[30] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[31] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[32] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022.

A Full set of LDA topics

Table 4 contains the full set of topics for the $k = 30$ LDA model introduced in § 4.

B Dataset Card

Our dataset card is available at `https://github.com/allenai/mmc4/blob/main/DATASET_CARD.md`

C Demonstrative Examples

C.1 Images w/ Watermarks/Ads/Logos

Figure 7a depicts a few sample images containing watermarks in various forms, Figure 7b shows images that are associated with logos, and Figure 7c lists a few sample images related to advertisements. Notice that the dissimilarity between images associated with logos and those pertaining to advertisements is relatively modest. Although images connected to advertisements may occasionally encompass promotional language or persuasive expressions, they may also solely feature logos. Notably, the principal criterion for determining whether an image is ad-related is contingent upon assessing its relevance to the document. If the image is less related to the document, it is more aptly categorized as ad-related. For instance, the interleaved document presented in Table 5 contains two images associated with logos that are intricately linked to the commercial brand being presented within the document. Consequently, these two images are not classified as advertisements.

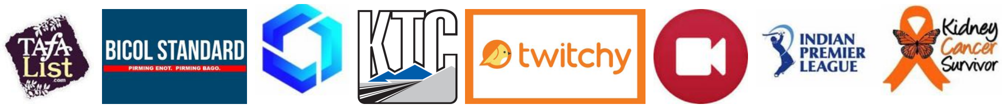
Topic name	Rate	Top Words
E-commerce	4.61%	products, quality, price, product, online, offer, buy, customers, services, order
Healthcare	2.55%	health, care, body, patients, treatment, medical, pain, cancer, blood, mental
Travel	3.98%	city, hotel, park, visit, travel, trip, tour, enjoy, beach, town
Celebrations	3.94%	fun, wedding, beautiful, christmas, happy, card, birthday, gift, blog, perfect
Music	2.50%	music, band, album, song, sound, songs, dance, show, live, musical
Religion	2.05%	god, church, jesus, lord, faith, man, father, heart, christ, gods
Fashion	4.86%	black, white, size, color, design, wear, style, fabric, cut, fit
Nature	3.05%	water, dog, river, fish, dogs, species, animals, fishing, sea, weather
Geography	3.56%	city, county, state, york, san, north, west, st, john, south
Business	4.15%	management, company, marketing, technology, data, services, team, industry, project, clients
Technology	4.89%	page, app, site, download, website, data, click, google, web, email
Education	2.39%	students, school, learning, skills, children, education, learn, student, training, class
Research	1.43%	data, download, research, analysis, study, al, cells, memory, studies, results
Food	3.31%	food, add, recipe, minutes, chocolate, cream, delicious, chicken, sugar, cheese
Law	2.14%	law, insurance, court, legal, case, state, letter, act, cover, policy
Wellness	1.92%	skin, hair, oil, natural, organic, wine, plant, products, plants, water
Self-improvement	5.27%	change, youre, mind, point, means, fact, thing, ways, question, process
Politics	2.73%	government, president, police, political, war, trump, military, state, party, security
Engineering	2.81%	water, energy, system, power, air, temperature, heat, systems, gas, solar
Sports	3.01%	game, games, team, play, season, players, win, league, player, football
Economy	2.29%	percent, market, million, —, trade, billion, growth, price, company, report
Architecture	3.08%	room, space, house, kitchen, floor, living, pool, building, large, bedroom
Automotive	3.20%	car, vehicle, camera, engine, power, system, model, control, speed, phone
Community	3.91%	community, university, program, research, members, support, development, public, national, group
Finance	1.72%	money, credit, card, real, property, estate, loan, pay, financial, tax
International	2.31%	international, india, countries, china, south, history, united, country, europe, indian
Events	3.93%	2018, event, pm, 2019, 2017, april, 2016, posted, friday, june
Literature	3.73%	book, story, books, film, series, movie, read, characters, stories, reading
Personal	7.96%	ive, didnt, thing, bit, thought, week, wanted, started, pretty, id
Art	2.70%	art, design, de, images, ikea, image, painting, collection, piano, photo

Table 4: LDA[6] topic modeling outputs (k=30 topics) when trained on a random sample of documents from `mmc4`. Topic frequencies are determined by taking the mean distribution over documents in the corpus. Topic names are generated by GPT-4 conditioned on the top 20 words for each topic, prompted by a request for a short 1-2 word summary.

C.2 Interleaved Document

Table 5 and Table 6 show two interleaved docs from `mmc4`, displaying the list of sentences and the corresponding assigned images, alongside the CLIP ViT/L-14 image-text similarity score.

(a) Images with watermarks.



(b) Images related to logos.

(c) Images related to ads.

Figure 7: Manually labeled images with watermarks and images related to logos or ads.

Sentence	Image	CLIP Similarity
`Our new service for teams to manage their fleets for racing.`		
`Getting boats has never been this easy.`		
`Get a step ahead with the planning for your team and get all the boats you need for next season races.`		23.51
`Our new service for teams to manage their fleets for racing.`		22.40
`As easy as adding boats to a list, this service aims to be the simplest way to rent boats, no extra knowledge needed and with full support from our staff.`		
`Get all the features of a Nelo boat, from having great equipment to our service team for a fraction of the price of a new boat.`	NELO	28.76
`All our rental boats for racing are carefully maintained and revised between each race so each boat is as good as new.`		

Table 5: An example document from mmc4 with interleaved sentences and images, together with the CLIP ViT/-14 image-text similarities. This document contains two logo-related images (the 2nd & 3rd images with "NELO") that are relevant to the content of this document, and are therefore excluded from the category of advertisement.

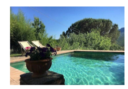
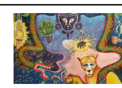
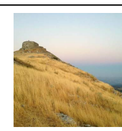
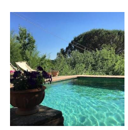
Sentence	Image	CLIP Similarity
`Are you thinking about running a retreat for your own group of people?`		25.93
`We are happy to help you hosting and organizing your own retreat.`		19.71
`We work with your interest in mind in designing your retreat, and we facilitate the logistics, supporting you all the way for a great experience.`		21.29
`Nestled within powerful and deeply inspiring nature, in the heart of Tuscany, Italy, Podere Di Maggio is a place born of dreams.`		22.35
`The dream to be close to and learn from nature.`		19.37
`The dream to create and share beauty.`		19.16
`The dream to discover and develop the poetry of being and doing.`		18.21
`We offer an invitation to explore a wide range of life arts: poetry, dance, music, yoga, meditation, ritual, ceramics, painting, singing, photography, seeing, hearing, touching, feeling, cooking, communicating and collaborating; sharing and daring to discover and unfold yourself.`		22.69

Table 6: A document instance retrieved from the `mmc4` dataset is presented, consisting of interleaved textual sentences and accompanying images, along with the CLIP ViT/-14 image-text similarity scores.