Learning Visual Composition through Improved Semantic Guidance

Austin Stone[*] **Hagen Soltau**[*] **Robert Geirhos** **Xi Yi** **Ye Xia** **Bingyi Cao**
Kaifeng Chen **Abhijit Ogale** **Jonathon Shlens**

Google DeepMind

Abstract

Visual imagery does not consist of solitary objects, but instead reflects the composition of a multitude of fluid concepts. While there have been great advances in visual representation learning, such advances have focused on building better representations for a small number of discrete objects bereft of an understanding of how these objects are interacting. One can observe this limitation in representations learned through captions or contrastive learning – where the learned model treats an image essentially as a bag of words. Several works have attempted to address this limitation through the development of bespoke architectures. In this work, we focus on simple and scalable approaches. In particular, we demonstrate that by improving weakly labeled data, i.e. captions, we can vastly improve the performance of standard contrastive learning approaches. Previous CLIP models achieved near chance rate on challenging tasks probing compositional learning. However, our simple approach boosts performance of CLIP substantially and achieves state of the art results on compositional benchmarks such as ARO and SugarCrepe. Furthermore, we showcase our results on a relatively new captioning benchmark derived from DOCCI. We demonstrate through a series of ablations that a standard CLIP model trained with enhanced data may demonstrate impressive performance on image retrieval tasks.

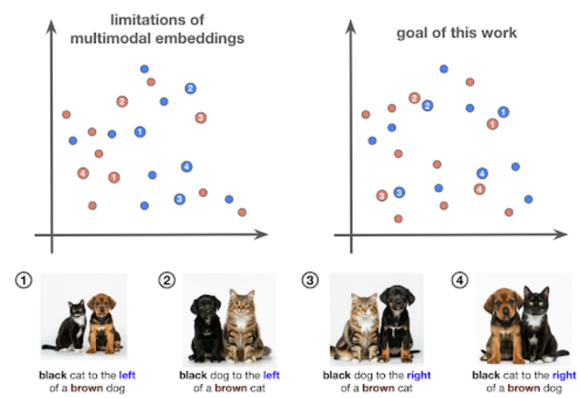

Figure 1. **Summary of results.** Left: Previous state-of-the-art results in multimodal embeddings have limited understanding of the composition of images [14, 26]. Right: The goal of this work is to learn multimodal embeddings which reflect a strong understanding of the composition of visual and semantic information. Images and captions are red and blue, respectively.

1. Introduction

Visual composition is a critical problem. Learning representations that capture how attributes and relationships of objects are represented is a critical task for any system that attempts to summarize the multitudes of concepts imbued within pixel space.

Many SOTA multimodal models do not learn embeddings that have an understanding of the relationship of objects and descriptions [22, 36]. One reason for this failure is that the visual representations are not rich enough to capture the interactions and relationships between various parts of an image or video.

Current approaches have focused on building new bespoke architectures that leverage multi-task learning through multiple losses or side data [17, 37]. Although these models achieve SOTA results on visual composition tasks [36], it is unclear how scalable these approaches are given limitations on gathering high quality, specialized side information or the complexity of the architectures.

Instead, we search for scalable solutions that focus on simplicity. The goal of this work is to employ a basic multimodal model solely employing a single contrastive learning objective [14, 26], and attempt to improve the learned representations with minimal changes.

We posit that underlying visual architecture (e.g. ViT [8]) contains sufficient parameters and scale to capture visual composition. Furthermore, we posit that contrastive

[*] Equal contribution.
Correspondence: {austinstone,soltau}@google.com

learning is a sufficient training signal for multimodal learning so long as the target semantic embedding is rich enough. In particular, we ask if enriching the target semantic embedding is entirely sufficient to make a contrastive learning model capture visual composition.

We answer this question by identifying a small set of minimal changes to improve the target semantic embedding of a model. First, we recaption our training data using a strong multimodal foundation model (Figure 2. Second, we replace the text tower for CLIP [14, 26] with a powerful text-based foundation model. We find that these two changes are entirely sufficient to greatly improve the visual embedding representation. We measure this effect using standard benchmarks [36] and discuss more newer benchmarks [22, 28] to showcase these performance gains. Notably, we find that for detailed image retrieval tasks, our approach improves the open-source CLIP model [26] recall @1 from 58.4% to 94.5%.

In summary, we focus on building simple and scalable techniques for improving compositionality in learned multimodal models focused on the problem of retrieval. We summarize our main contributions as follows:

- Rich visual captions may be automatically generated using foundational models with simple techniques to minimize hallucinations (Sec. 2, 3.3, and 3.5).
- Rich visual captions are sufficient for correcting many of the failures in compositional representation (Sec. 3.1).
- COCO captioning benchmark [19] for measuring visual representations is saturated and subject to overfitting (Sec 3.2). New benchmarks are necessary for measuring improvements [22, 36] (Sec 3.3).

2. Methods

Our method builds on the CLIP architecture proposed in [13, 26]. We introduce two key modifications:

2.1. Semantic guidance with grounded recaptioning

We start with an English-only subset of the WebLI dataset [4, 5] consisting of 1B high quality images paired with freeform text scraped from the corresponding `alt-text` within a web page. We posit that the original `alt-text` is noisy and a limiting factor in the performance of an multimodal model [16]. Thus, we focus on creating a new set of captions to improve the `alt-text`.

To enhance caption quality, we leverage Gemini 1.5 Flash [27] to generate new captions. In early experiment, we provide the original image as well as the `alt-text` and web page title, and prompt the model to generate a new caption that describes the underlying image. We iterated on the prompt (see Appendix for details) to arrive at a method that minimizes hallucinations while providing rich descriptions.

Figure 2 summarizes our simple procedure. These examples showcase features in the recaptioning procedure. First, the grounding text supplied by the web page title and `alt-text` provides rich grounding information which can be exploited by the model. Second, the multimodal model can employ OCR in the original image to improve the noisy `alt-text` and correct for errors.

In early experiments, we additionally found significant gains by providing two forms of data augmentation. First, we performed sentence sampling, where by randomly selected subsets of caption sentences are used as targets for the model. Second we synthesized 2 million "hard negative" examples to our training mixture (Figure 3). We explored the choices of these two data augmentations in Section 3.5.

The resulting captions contain a mean of 57 words, increasing the effective caption length by about $8\times$ of the `alt-text` (Figure 4) and likewise significantly increasing the log-likelihood of the caption as measured by Gemini Pro 1.5 (Figure 5). In our ablations we experiment with an assortment of other captioning techniques. We generate a "concise" captions (26 $\pm$17 words); we examine the importance of grounding the caption based on the `alt-text` and page title; finally, we experiment with using a weaker LLM's as the captioning model (Section 5 and see Appendix for additional details).

2.2. Semantic guidance with a strong text encoder

Instead of training a text encoder from scratch, we utilize the pretrained Gemini 1.5 Flash-8B [30] as the text encoder. We experiment with unfreezing various layers to allow for more flexibility in training the visual embedding. In early experiments, we explored unfreezing various layers of these pretrained text encoders and found that unfreezing last 4 layers of this model provides best overall performance relative to the additional compute cost during training. We repeated the above experiments with the open source Gemma2-2B [31] as the text encoder, and likewise identified that unfreezing the last 4 layers provides the best trade off in terms of additional compute cost versus performance. This model is smaller than the Gemini model above, but yields comparable results.

3. Results

Our model is a standard two-tower CLIP encoder trained with a contrastive loss [26]. The image encoder is a ViT-Base vision transformer with 86M parameters [8]. Images are cropped to 256×256 pixels and represented as 256 patches of 16×16 pixels. The encoders represent images and text as 768-dimensional embeddings.

For the text encoder, we experiment with two different configurations: In the first configuration, a bi-directional text encoder with 12 layers and 297M parameters is trained from scratch matching CoCa-Base [35]. We explore also

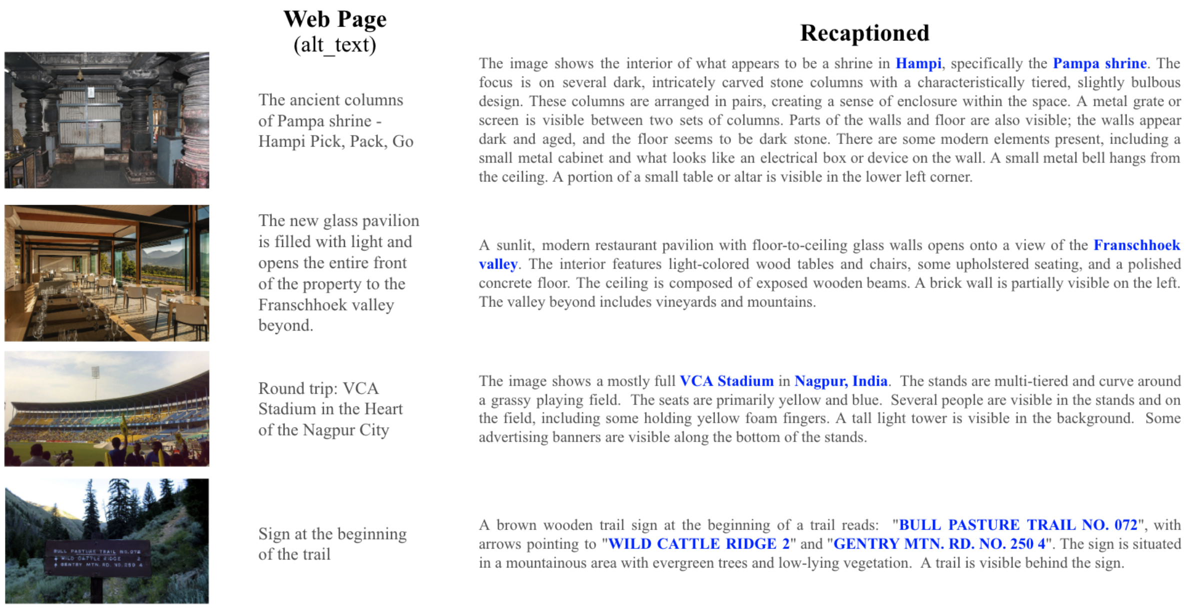

Figure 2. **Summary of methodology.** All images were recaptioned using a multimodal foundation model grounded on the image and the `alt-text` for the image on the web page. In this case, we show captions generated using Gemini 1.5 Flash [30]. We highlight aspects of the new caption in **blue** that leverage the `alt-text` or demonstrate a capability. Note how the generated captions leverage information provided by the `alt-text` or perform OCR on the original image to improve the caption. Captions are for images from CC-12M [3].

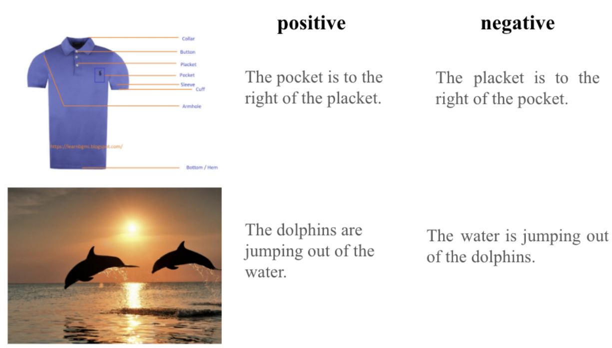

Figure 3. **Synthetic negative captions.** We prompt a foundation model [30] to generate 64 million synthetic positive and negative annotations. To generate the negative prompts, we provide few shot examples matching the style of ARO relations and attributes evaluation. Captions are for images from CC-12M [3].

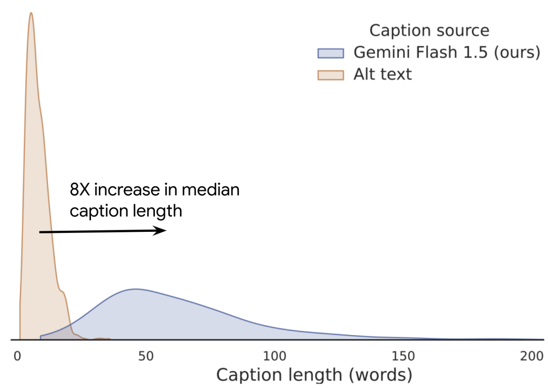

Figure 4. **Recaptioning increases caption length by 8X.** While the median alt text caption length is just 7 words, the detailed captions generated by Gemini Flash 1.5 increase this to 57 words.

the use of a different text encoder based on a pre-trained Gemini 1.5 Flash-8B [1] and Gemma2-2B [31]. We keep most of the layers frozen, and only fine-tune the last 4 layers. While the pre-trained text encoder uses left-to-right context only, we switch to bi-directional context for the layers that are finetuned. In total, the model has 653M trainable parameters. We perform 150,000 training steps with global batch size of 65,536 followed by fine-tuning for 500 training steps with reduced batch size of 4096 using data augmentation on the "hard negative" examples (Section 2.1). The main model training is done on 256 accelerators and the fine-tuning is done with 16 accelerators. We use Adam optimizer [15] with a linear warm-up of the learning rate.

3.1. Beyond a bag of words representation

A primary indication of the failure of multimodal representations can be observed a the relatively simple ARO [36] and SugarCrepe [11] evaluation datasets. ARO artificially perturbs a set of captions from "*the horse is eating grass*"

[1] `https://ai.google.dev/gemini-api/docs/models/experimental-models`

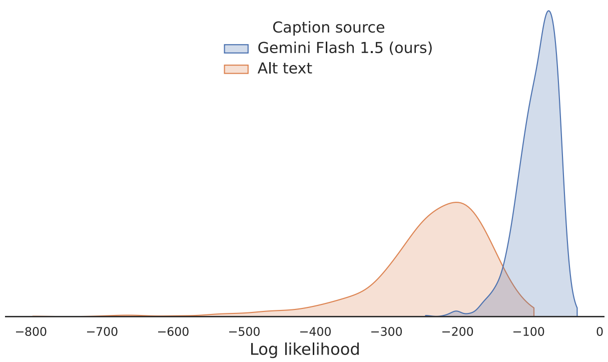

Figure 5. **Recaptioning improves caption log-likelihood.** Alt text on the web is often unnatural (example: "`bigtimerush nyc 007`"), leading to low log-likelihood with a median of -223. In contrast, the captions from Gemini Flash 1.5 substantially improve median log-likelihood to -83, indicating that these captions are a lot closer to natural language and sentences than alt text.

	relations	attributes
Gemini-text [30]	71%	82%
Vera [20]	62%	83%
CLIP [26]	59%	63%
CoCa [35]	48%	50%
NegCLIP [36]	71%	81%
BLIP † [17]	59%	88%
X-VLM † [37]	73%	87%
Ours	92%	94%

Table 1. **Our model improves upon bag of words.** Performance on Attribution, Relation, and Order (ARO) benchmark [36]. All numbers report classification macro accuracy. Chance rate is 50%. Note that BLIP and X-VLM employ a second-stage binary classification in order to exhaustively identify the best candidate. Vera and Gemini-text only see the caption and not the images.

to "*the grass is eating a horse*" (i.e. relations) or "*the paved road and the white house*" to "*the paved house and the white road*" (i..e. attributes), and subsequently asks if the corresponding image is closer to the former or the latter caption. Chance rate is 50% and a model that does respect the relations and attributes of an image – such as a bag-of-words – would correspondingly perform at chance rate. One of the most widely used multimodal baselines, CLIP [26], achieves 59% and 63% accuracy on relations and attributes dimensions of ARO, indicating that the model is performing only slightly above a bag-of-words representation (Table 1).

Admittedly, this task is artificial, and thus some captions can be correctly identified without examining an image (e.g., "*grass cannot eat a horse*") merely based on the language. To show this, we supplied Gemini 1.5 with two caption alternatives but *no* image. This *blind* baseline achieves an accuracy of 71% and 82% on relation and attributes, respectively, outperforming CLIP. SugarCrepe improves upon this deficiency such that even text-only models specifically trained for verbal plausibility such as [20] only score at random chance levels. Apart from the more plausible negative captions, SugarCrepe is similar to ARO. We additionally show our results on SugarCrepe in Table 2.

	replace			swap		add	
	O	A	R	O	A	O	A
Human [11]	100%	99%	97%	99%	100%	99%	99%
Vera [20]	49%	50%	49%	49%	49%	49%	50%
CLIP [13]	94%	79%	65%	60%	62%	78%	72%
DC-XL [10]	96%	85%	70%	65%	67%	91%	85%
LAION [29]	97%	86%	72%	64%	72%	93%	86%
CapPa [32]	92%	90%	87%	82%	88%	99%	99%
GPT-4V [23] †	96%	94%	90%	83%	90%	92%	92%
Ours	97%	94%	88%	89%	94%	95%	93%

Table 2. **Results on SugarCrepe [11]** O, A, and R stand for the object, attribute, and relation split, respectively. Chance rate is 50%. Vera is a text-only model which does not see the images. † GPT-4V sees both captions simultaneously and is not a two tower embedding model. LAION is `xlm-roberta-large-ViT-H-14`.

Several directions have been pursued for improving performance on these baselines (Table 1, 2). The ARO benchmark proposed a training augmentation method NegCLIP in which hard negatives are artificially added to the training set, which achieves 81% and 71% accuracy. More importantly, several works have built bespoke architectures, derived from CLIP, which exploit localization information in order to better ground visual information. Correspondingly, two different architectures achieves SOTA performance on relations (73% [37]) and attributes (88% [17]). Both models achieve these results by tieing the image and text towers together, using a grounded, cross-modal encoder [17, 37]. Although both methods employ a standard nearest neighbor lookup, they achieve SOTA results through a second-stage computation that exhaustively calculates a binary classification score for a set of candidate image-caption pairs. [32] replaces the contrastive objective with a captioning objective and shows this leads to gains on vision & language tasks.

In comparison, our work makes no architectural changes and maintains a single contrastive training objective in a standard CLIP model. We do employ additional data augmentation with 64M synthetically generated "hard negative" examples (Section 2.1) and find that the addition of these data augmentations significantly improve compositional understanding (Section 3.5). Notably, our model achieves 92% and 94% accuracy on relations and attributes surpassing bespoke architectures that exhaustively search across examples. On SugarCrepe, we achieve state of the art or highly competitive results across results across all splits, even surpassing GPT-4V [23] on all but one split. In the Appendix, we measure how our model uniformly outperforms competing methods across the assortment of fine-grained tasks comprising the ARO benchmark.

3.2. Limitations of retrieval with COCO captions

Given the positive results on ARO, we asked whether these gains can be observed in image retrieval on the COCO dataset [19]. COCO contains detailed captions for 5,000 images and we ask how well nearest neighbor lookup in the embedding space performs for image retrieval. Our results compared to external baselines are in Table 3.

As a strong baseline, we measure the performance of open source CLIP [26] and a CoCa [35] on image retrieval. We find a recall @1 of 37.8 and 47.5 on for CLIP and CoCa, respectively. Our model achieves the best reported recall of all models of similar size with a recall @ 1 of 56.3.

Although our result was stronger than prior numbers, we actually expected to see even more notable gains in our model given the high quality and scale of our data. We investigated our model further by examining cases where our model is marked as retrieving the incorrect COCO image. Figure 6 showcases three randomly selected examples. Note that although these examples are marked as incorrect, the image retrievals looks quite plausible. In fact, upon further inspection of the dataset, we observe that many COCO captions contain similar images and captions, and in the context of image retrieval, we hypothesize that many image retrieval errors are *incorrectly* scored as errors.

To test this hypothesis, we did a human annotation experiment to assess what fraction of images retrieved by our model are reasonable. We randomly select 500 failure cases and task 3 humans to annotate each of the failures to indicate whether the recalled image is an appropriate match to the prompt text. [2] Using majority voting among the 3 human raters, we measured that 29.8% of the image retrieval failures were appropriately marked failings of our model, but the other 70.2% were acceptable matches to the query text. In other words, 70.2% of all instances marked as failures were images that a majority of the human raters thought were appropriate image retrievals given the input caption.

These results indicate that a majority of the errors reported in Table 3 are due to noise and overly similar images or annotations in COCO when employed as an image retrieval task. This result suggests that caution must be exercised when fine-tuning and evaluating on COCO, and motivates our interest to identify another evaluation benchmark that might provide a useful metric for measuring improvements in image retrieval that exercises a detailed semantic representation.

	COCO	DOCCI-test	DOCCI-full
CLIP [26]	37.8	58.4	46.2
CoCa [35]	47.5	59.1	53.2
BLIP [17]	–	54.1	–
X-VLM [37]	56.1	–	–
VeCLIP [16]	48.9	–	–
Long-CLIP [40]	40.4	45.2	–
MATE [12]	–	73.4	–
TIPS [2]	59.4	58.8	–
Ours	56.3	94.5	89.7

Table 3. **Image retrieval based on caption** on COCO [19] (5,000 images), DOCCI [22] test split (5000 images) and the entire DOCCI dataset (15,847 images). All numbers report recall @ 1. CoCa [35] is the base model. All models were *not* fine-tuned on COCO or DOCCI. DOCCI results for CLIP and CoCa reproduced from checkpoints. DOCCI results for BLIP and Long-CLIP reported in [12].

3.3. Evaluation on detailed image caption retrieval

Given the positive results on the ARO benchmark but the inability to further discern gains on COCO, we investigated whether another retrieval benchmark may capture the gains in our approach. The DOCCI dataset was recently introduced in the context of image generation [22] but has recently been explored for image retrieval. DOCCI contains 14,847 images with detailed *human-written* descriptions (see [22] for details). Briefly, each caption contains on average 135.9 words across 7.1 sentences . Importantly, DOCCI purposefully contains images and corresponding captions in which there exist slight changes and alterations to the pose, orientation, background, or action. We construct an image retrieval task from this dataset and measure recall @ 1 for image retrieval. We report results on this dataset primarily in Table 3, but see Tables 5, 6, 7, and 8 for detailed ablations.

Open-source CLIP baseline achieves 58.4% recall. Notably, recently reported results on newer bespoke architectures with additional losses and side information achieve up to 73.4% recall @ 1. The best performing model employs a novel projection module to better align text and visual representations [12]. The authors additionally fine-tune on DOCCI train split at an increased image resolution of 448×448 pixels to achieve their best result of 84.6% recall @ 1. We find that our model achieves 89.7% recall @ 1 in a zero-shot manner using a standard CLIP training model

[2] During experimentation, we noticed that a notable fraction of the failure cases were because the ground truth caption for a given image was incorrect (i.e. in some examples, there is no image in the dataset which matches a given caption). We only assessed whether the recalled image matches the prompt caption, but this number might under report the true false failure rate due to some prompts having no appropriate image for our model to recall.

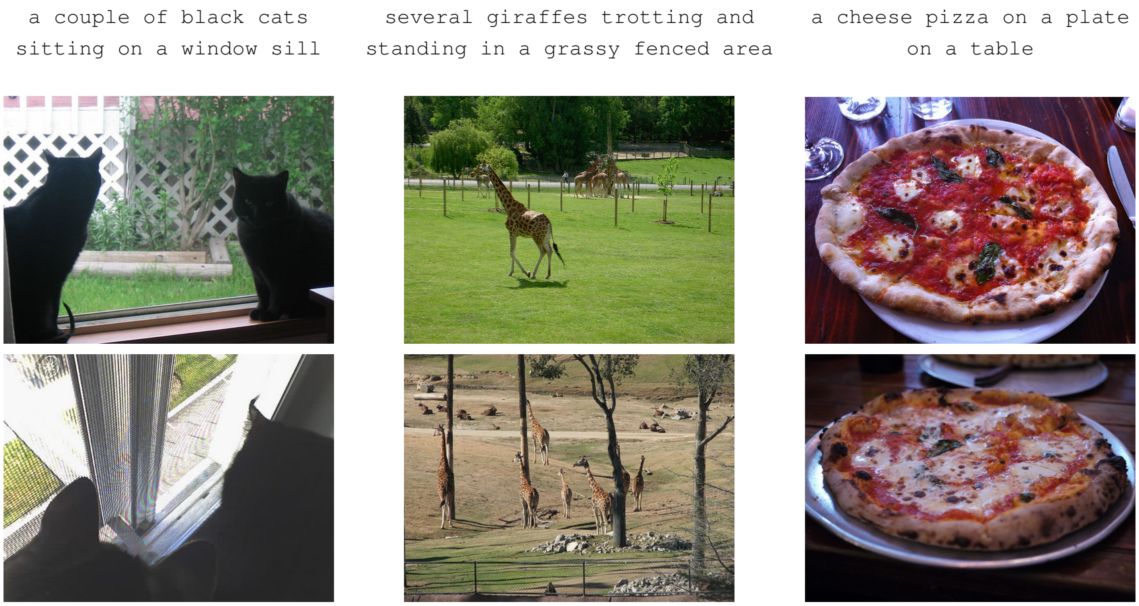

Figure 6. **Qualitative examples of limitations of image retrieval COCO captions.** Three examples showcasing where the model recalled image was incorrect, but is well described by the caption. **Top and Middle Rows**: Caption and associated ground truth image. **Bottom Row**: Top recalled image from our model. Note that the recalled images are well described by the caption corresponding to the original ground truth image.

but with our improved text supervision (but see Section 3.5 for even higher numbers). We note that our model does not fine-tune on DOCCI train split, and uses a reduced resolution of 256×256. We would expect that employing both well-known techniques, e.g. boosting image resolution and fine-tuning, would yield higher performance as well, but reserve such experiments for future work.

3.4. Evaluation on zero-shot ImageNet classification

Prior work on embeddings and image retrieval have additionally evaluated on ImageNet and variants [6]. We emphasize that ImageNet emphasizes centrally-cropped, single objects within a scene, and *not* the composition of multiple ideas and objects. Nonetheless, this dataset has proven to be a standard, reported metric in computer vision and, thus we apply our model to this task. We find that our model achieves a zero-shot performance of 68.4% top-1 accuracy (Table 4). Although a reasonable zero-shot performance, we find nonetheless that this result is below other SOTA zero-shot ImageNet classification accuracies (Table 4).

One immediate discrepancy between our work and prior work is the source of training data. The open-source CLIP model employs a separate, proprietary training set that has never been revealed, so it is unclear how well this aligns with ImageNet classification task [3]. CoCa [35] (and other multimodal embedding models, e.g. LiT [39], BASIC [25]) leverage a 50% mixture of variations of the JFT dataset. In the case of JFT, the dataset comprises 3 billion images that have been annotated with a detailed class-hierarchy across 30K labels using a semi-automatic pipeline [38]. We strongly suspect this fine-grained information contributes heavily to their overall performance.

To validate this hypothesis, we likewise introduce a 50% mixture of JFT to our training set, and achieve 79.1% zero-shot performance, suggesting that indeed the distribution of JFT is well tailored to ImageNet. Although our result is still below the best reported zero-shot result of 82.6%, we emphasize that there are a host of techniques that one could exploit to further boost this performance – and were specifically employed by [26, 35] [4]. We decided not to pursue this direction, and instead view the ImageNet results as demonstrating the model can indeed be performant for fine-grained discrimination if supplied the appropriate data. However, we view the topic of fine-grained classification as somewhat orthogonal to the central problem of building a model that better handles visual composition.

[3]We observe that CLIP achieves statistically significantly higher accuracy on a class-balanced version of the ImageNet training set than it does on the ImageNet validation set (77.5% train vs 76.2% validation, two-proportion z-score$= 6.2$, $p < 1 \times 10^{-9}$), indicating that some training data contamination maybe occurring for this model.

[4]Two well-known techniques include fine-tuning models on high resolution data, and explicitly matching the captioning text used in ImageNet zero-shot evaluation in the training set. Both techniques provide additive gains for boosting performance [35]

	ImageNet ZS accuracy
CLIP [26]	76.2
ALIGN [13]	76.4
CoCa [35]	82.6
Ours	68.4
Ours (with JFT [38])	79.1

Table 4. **Zero-shot image classification results on ImageNet**. Prior work performance from [35]. CoCa report base model.

Text Encoder	Text data	COCO	DOCCI-full
Scratch	`alt-text`	47.8	53.5
Gemini-8B	`alt-text`	48.3	67.2
Scratch	re-captioned	46.5	75.6
Gemini-8B	re-captioned	51.9	91.6
Gemma-2B	re-captioned	51.2	88.9

Table 5. **The importance of pretraining the text encoder and recaptioning.** We experimented with both Gemma [31] and Gemini [30] based pre-trained text encoders and see similar performance. Results were collected without hard-negative fine-tuning or random sentence sampling.

	COCO	DOCCI-full
Gemini 1.5 Flash [30]	39.2	90.3
Gemini 1.5 Flash-8B [30]	37.5	88.4

length	grounded?	COCO	DOCCI-full
default	✓	39.2	90.3
default		31.5	89.3
short	✓	39.0	85.2
alt-text	✓	38.4	65.7

Table 6. **Exploration of large-scale captioning** measured with recall @ 1 on image retrieval on COCO [19] and DOCCI [22]. All comparisons based on 100M captions from internal WebLI dataset [4]. **Top**: Comparisons across different models for recaptioning. **Bottom**: Comparison based on altering method for recaptioning. Grounded indicates whether `alt-text` is supplied to the caption generation. Default and short caption length contain 133.4 and 354.8 words, respectively. Default captions used for rest of paper. Results were collected without hard-negative fine-tuning or random sentence sampling.

3.5. Exploring the space of improved guidance

We examine the space of semantic guidance to better understand the additive benefits of each change in our model. Given that captioning 1B images is prohibitively expensive, we perform detailed ablations on a subset of 100M images for image retrieval on DOCCI and COCO (Tables 5, 6, 7, and 8).

Table 5 indicates that training on recaptioning data alone notably boosts performance from 53.5 to 75.6 for recall@1 on DOCCI [22]. We emphasize that the training images are identical; all that differs is the associated captions with the images. This small change results in 41% relative performance gain. In parallel, we observe that using a rich text model as a target increases performance from 53.5 to 67.2, corresponding to 26% relative performance boost. The two simple changes in tandem lead to additive gains on COCO and DOCCI.

Given the magnitude of benefits with recaptioning, we explored what features make for a good recaptioning. First, we examined how a less performant recaptioning models effects overall performance (Table 6, top). We observed that slightly drop in performance with a less performant recaptioning model. Second, we examined how two key parameters in our recaptioning with Gemini 1.5 Flash effect overall quality. Namely, we removed grounding based on `alt-text`, and we abbreviated length of the captions. We likewise observed that only slight degradations in overall performance on image retrieval (Table 6, bottom). Finally, we asked how data augmentations for the captioning effected performance. In particular, we checked the impact from random sentence sampling (Table 7) and hard negatives fine-tuning (Table 8). We observed that sentence sampling uniformly lead to better performance, especially on COCO. Conversely, removing the hard negatives did positively impact DOCCI at the significant harm to ARO and minor harm to COCO. We expect more work should be dedicated to finding the right balance of data augmentation in captioning pipelines in future work.

Sentence Sampling	COCO	DOCCI-full
	51.9	91.6
✓	56.3	93.0

Table 7. **Impact of random sentence sampling.** We randomly sample between one and ten sentences from our long generated captions. Doing so boosts recall on COCO and DOCCI. Results were collected without hard-negative fine-tuning.

Hard-neg. FT	COCO	DOCCI-full	ARO Attr.	ARO rel.
	51.9	91.6	82%	65%
✓	54.1	88.1	94%	93%

Table 8. **Impact of hard-negative fine-tuning.** We see that hard-negative fine-tuning significantly boosts results on ARO, slightly helps COCO, and slightly harms DOCCI. Results were collected without random sentence sampling.

4. Related Work

Shortcomings of multimodal models. Since the landmark publication of CLIP [26] as one of the first highly capable vision-language models, a number of important limitations have been observed in multimodal models. For instance, many image pairs are highly similar in CLIP feature space despite showing semantically meaningful differences (e.g., left-facing dog vs. right-facing dog). This leads to systematic failures in downstream models that use CLIP as an image encoder (although [18] observed that better image-language alignment can sometimes alleviate this concern). Furthermore, vision-language models often behave like bag-of-word models [36], as indicated through a surprising inability to understand compositions and relations (e.g., grass eating horse vs. horse eating grass). A number of approaches have been suggested for improving over current shortcomings; they are mostly focused on specialized architecture or data, and sometimes both.

Specialized architectures. X-VLM [37] proposes an architecture that incorporates bounding box prediction in order to improve the connection between visual input and text descriptions. The BLIP model [17] is a popular specialized architecture trained jointly with three different objectives. CapPa [32] suggests image captioning as an alternative pretraining task leading to improved downstream performance. MATE [12] identifies short captions as a problem (similar to our work); however instead of training models on longer captions (as we do) they propose an architecture that replaces the text encoder with a LLM-based one. Long-CLIP [40] has a similar motivation and uses a bespoke architecture (modifying positional embedding, matching CLIP's primary component) in combination with one million long-text caption pairs. [2] is a concurrent work which uses multiple embedding heads, one for fine-grained tasks like ImageNet and the other for long form captions like DOCCI, in addition to a very large pretrained DINO g/14 image encoder and additional self-distillation and masking losses.

Data. In contrast to the wealth of specialized architectures, the data side has not been explored as much—and frequently only in combination with a specialized architecture. Among the notable exceptions are [24], generating synthetic images from text-to-image models that are used as hard negatives in contrastive training. Furthermore, [9] use an LLM to rewrite (but not explicitly expand) existing captions as a form of data augmentation. In contrast to our work, the caption rewriting model is not visually grounded since it is a "blind" language model without image access. Finally, closest in spirit, [16] employed a multimodal model to more elaborate recaptioning pipeline but focused on boosting COCO retrieval.

5. Discussion

In this work we have highlighted the failures of multimodal embeddings. Current foundational models contain limited compositional information as observed by perturbations in the attributes and relationships as well as the failure of models to perform image retrieval on challenging captioning tasks (Section 3.1, 3.2, and 3.3).

This work attempts address this problem by providing improved semantic guidance for multimodal models. In lieu of architectural innovations, this focuses on two simple and small changes that are highly scalable: (1) re-captioning training, and (2) improving the semantic target with a pretrained and powerful language model. These two improvements appear complimentary and result in a model that is able to perform 94.5 recall @1 on DOCCI image retrieval [22] (Table 3).

More generally, this work highlights how bootstrapping data provides a scalable solution for constantly improving training data. Previous work had shown how simple procedures to bootstrap pseudo-labels may scale quite well and achieve state-of-the-art results on discrete classification tasks [34]. This work extends that endeavor by using pseudo-labels from other models to learn better embedding representations using contrastive learning.

As a simple application of this work, one benefit of improved alignment between visual and semantic representations is the opportunity to predict human preferences for image generation models. For instance, in previous work Imagen [28] outperformed GLIDE [21] in forced-choice human preference experiments for whether a caption is better aligned to the generated image. As a pilot study, we asked whether our work may improve upon a strong baseline of the open-source CLIP model [26] for predicting these human preferences on the same dataset. We observed that while open-source CLIP predicted the weighted human accuracy at 71.3%, our model instead achieved 81.1% weighted accuracy. We also observed notable gains in terms of the ability of our model to count a small number of distinct objects. We take these preliminary experiments as a suggestion for a potentially fruitful direction of exploration and reserve such a direction for future work.

This work has largely measured the quality of the model based on image retrieval in a shared embedding space. That said, a natural extension of this work is to investigate how these simple techniques may work in the context of an image captioning model (e.g. [33]). Previous work has shown how image captioning as a training task may provide rich representation that may transfer to other visual tasks [7]. A natural question is to ask how such a training objective may additively improve multimodal representations [35]. In tandem, one could measure the quality of the presentation on captioning-based evaluations [1] where challenges such as grounding and hallucination become more pronounced. We look forward to pursuing this direction in future work.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. 8

[2] Anonymous. TIPS: Text-image pretraining with spatial awareness. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review. 5, 8, 12

[3] 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. 3

[4] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2, 7

[5] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 6

[7] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations, 2021. 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1, 2

[9] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites, 2023. 8

[10] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. 4

[11] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3, 4

[12] Young Kyun Jang, Junmo Kang, Yong Jae Lee, and Donghyun Kim. Mate: Meet at the embedding – connecting images with long texts, 2024. 5, 8, 12

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2, 4, 7

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. 1, 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 3

[16] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. Veclip: Improving clip training via visual-enriched captions, 2024. 2, 5, 8, 12

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 1, 4, 5, 8, 12

[18] Siting Li, Pang Wei Koh, and Simon Shaolei Du. On erroneous agreements of CLIP image embeddings. *arXiv preprint arXiv:2411.05195*, 2024. 8

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2, 5, 7

[20] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A general-purpose plausibility estimation model for commonsense statements, 2023. 4

[21] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 8

[22] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *arXiv:2404.19753*, 2024. 1, 2, 5, 7, 8

[23] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian,

Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 4, 5

[24] Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. *arXiv preprint arXiv:2411.02545*, 2024. 8

[25] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning, 2023. 6

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4, 5, 6, 7, 8, 12

[27] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 8

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 4

[30] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 2, 3, 4, 7, 12

[31] Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. 2, 3, 7

[32] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too, 2023. 4, 8

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016. 8

[34] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020. 8

[35] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 2, 4, 5, 6, 7, 8, 12

[36] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 1, 2, 3, 4, 8, 13

[37] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts, 2022. 1, 4, 5, 8, 12

[38] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, 2022. 6, 7

[39] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022. 6

[40] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip, 2024. 5, 8, 12