

I can't believe there's no images!

Learning Visual Tasks Using only Language Data

Sophia Gu* Christopher Clark* Aniruddha Kembhavi
 Allen Institute for Artificial Intelligence
 {sophiag, chrisc, anik}@allenai.org

Abstract

Many high-level skills that are required for computer vision tasks, such as parsing questions, comparing and contrasting semantics, and writing descriptions, are also required in other domains such as natural language processing. In this paper, we ask whether this makes it possible to learn those skills from text data and then use them to complete vision tasks without ever training on visual training data. Key to our approach is exploiting the joint embedding space of contrastively trained vision and language encoders. In practice, there can be systematic differences between embedding spaces for different modalities in contrastive models, and we analyze how these differences affect our approach and study a variety of strategies to mitigate this concern. We produce models using only text training data on three tasks: image captioning, visual entailment and visual question answering, and evaluate them on standard benchmarks using images. We find that this kind of transfer is possible and results in only a small drop in performance relative to models trained on images. We also showcase a variety of stylistic image captioning models that were trained using no image data and no human-curated language data, but instead text data from books, the web, or language models. Our code is available at <https://github.com/allenai/close>.

1. Introduction

Although vision and natural language processing (NLP) tasks are typically thought of as being very distinct, there is often a high degree of overlap in the skills needed to complete them. Visual question answering and reading comprehension question answering both require parsing and understanding questions, visual entailment and textual entailment require comparing different semantic meanings, and captioning and summarization require writing text that summarizes the semantics of the input. This raises an intriguing

possibility: if a model learned to complete one of these tasks using a high-level semantic representation of the input text, then in theory it could immediately be able to complete the corresponding visual task as long as the input image is encoded in the same semantic representation. In this paper, we study whether this is possible by training models to complete a task using only natural language data, and then testing them on the same task with visual inputs instead of text. We call this setting *zero-shot cross-modal transfer* because it requires applying skills learned from one modality to a different one.

Accomplishing this requires encoding images and text into a shared semantic space. We use vision and language (V&L) models trained with a contrastive loss for this purpose [22, 47]. These models learn to embed text and images into vectors such that the vectors for matching images and captions are close together, and vectors for unrelated images and captions are far apart. Although this loss was originally intended for representation learning and zero-shot classification, here we show it also facilitates cross-modal transfer.

We propose a method called Cross modal transfer On Semantic Embeddings (CLOSE) to take advantage of these encoders. An outline of CLOSE is shown in Figure 1. During training, the text inputs are encoded into a vector using the (frozen) text encoder, which is then used as an input to a model. During testing, the visual input is embedded with an image encoder and used in place of the text embedding. Because these encoders were explicitly trained to produce embeddings that encode semantics in similar ways, learning to read and process the text vector should naturally translate to the ability to read and process the image vector. Although we focus on text-to-image transfer in this paper, our approach is applicable to other contrastive models such as videos [71], point clouds [1], and audio [10, 19, 69], potentially allowing transfer between many other modalities.

One potential difficulty with this approach is that, while contrastive embeddings do share some structure between modalities, there can still be significant differences between the image and text vectors in practice [35]. To mitigate this, we propose to additionally use *adapters* that modify the text

*Equal contribution

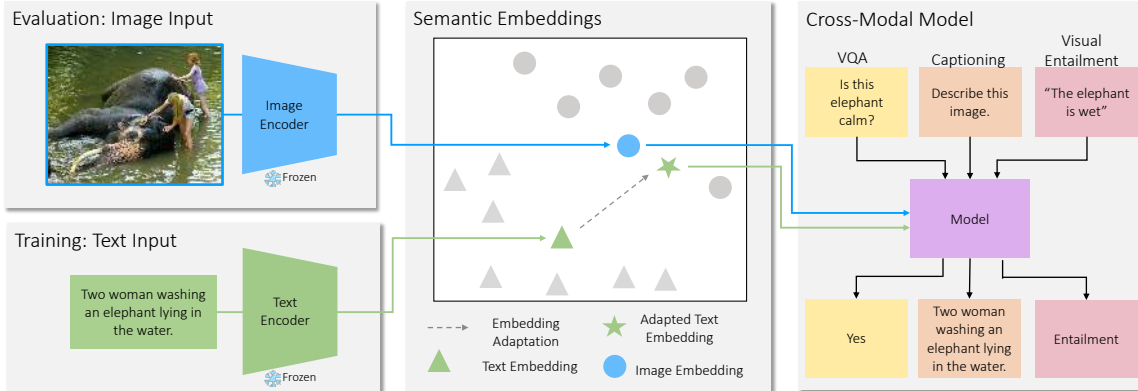


Figure 1. Overview of CLOSE. During training, input text is encoded into a vector with a text encoder which is then adapted with an adaptation method. A model learns to use the vector to perform a task such as VQA, captioning, or visual entailment. During testing, an input image is encoded with an image encoder instead to allow cross-modal transfer.

vectors being used during training. We find adding Gaussian noise to be very effective in boosting performance, but consider other approaches as well in our analyses.

We perform several experiments to better understand the effectiveness of our approach. First, we train models for the tasks of captioning, visual question answers (VQA), and visual entailment using only text data – typically using a caption describing a scene as a stand-in for an image, and then test the models on real images. We find these models only slightly underperform versions trained directly on images, demonstrating a high degree of transfer between the two modalities. Next, we show captioning models can also be trained using data generated from a language model, and therefore requiring almost no human-annotated data, and still acquire strong captioning competency. Finally, we complete two analyses: A sensitivity analysis showing that CLOSE is robust to cases where text and image vectors differ by a constant offset, which therefore allows CLOSE to work despite seemingly large differences between the image/text embeddings. Additionally, a study on the effectiveness of using an auxiliary vision and language corpus to build an improved adapter. We find that improvements are possible but vary depending on the source of that data, and that a particularly effective approach is to use the auxiliary data to compute a structured covariance matrix for use when adding Gaussian noise.

Achieving these results raises the possibility of training certain kinds of computer vision models using text data, which has some interesting potential applications. Text data is often easier to obtain than visual data because it can be directly constructed by annotators, or even by a large language model such as GPT-3 [3]. As a result gathering text to teach a model a particular skill can be significantly less expensive than gathering and annotating images for the same purpose. It also creates the possibility of directly using existing natural language datasets or other existing text

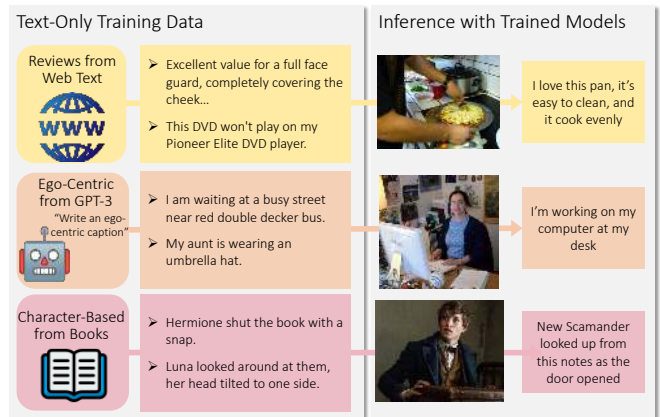


Figure 2. Using CLOSE to learn stylistic captioning without image data. Text examples of the desired style are gathered from sources such as the web, books, or GPT-3. Models are then trained on the text and then applied to images.

corpora to help train computer vision models.

We demonstrate one of these applications by training several stylistic captioning models using only text training data (see Figure 2). We collect text with various styles from a diverse set of sources, including internet reviews, books, and generations from GPT-3, and demonstrate that models trained on this text using CLOSE can produce accurate and stylistically correct captions for images.

2. Method

In this section, we discuss contrastive models and CLOSE in more detail.

2.1. Contrastive Learning

First, we provide a small amount of background on contrastive learning for V&L models. Contrastive learning



Figure 3. t-SNE [61] plots for various adapters on 350 image vectors (blue) and paired caption vectors (orange) from COCO captions. The first two panels demonstrate CLOSE, and the remaining three show additional adapters we study in our analysis (Section 4).

trains an image encoder and a text encoder to embed an image or sentence into a vector. These encoders are trained on the InfoNCE loss [60], where a batch of N matching images/captions are encoded to produce a set of image vectors, $\langle u_1, u_2, \dots, u_n \rangle$, and text vectors $\langle v_1, v_2, \dots, v_n \rangle$. For each pair of image/text vectors, a logit is computed as $l_{ij} = e^t \frac{v_i \cdot u_j}{\|v_i\| \|u_j\|}$, that is, as the cosine similarity between the vectors times a learned scaling factor, t .

These logits are then used in a cross-entropy loss with the logits of matching image/text vectors as positive examples, and the logits of the non-matching image/text vectors as negative examples. This trains the encoders to make the cosine similarity between paired image/text vectors large and the similarity between unpaired image/text vectors small, which inspires our approach of treating them as vectors in a shared semantic space. Existing implementations apply this method on hundreds of millions of images and have been shown to learn a wide range of visual concepts [21, 22, 47].

2.2. Model

Our model uses the image/text encoder from a contrastive model to encode the input, and then follows many prior works (e.g., [6, 24]) by fine-tuning a pre-trained language model to process this vector, along with any additional input text, to generate output text. First, the input image or text vector is normalized to have unit length to match what is used in the contrastive loss. Then that vector is converted into a number of vectors, we use 4 in our experiments, of the same dimensionality as the language model’s embedding layer using a linear layer. Next, other input text (e.g., the hypothesis in visual entailment or the question in VQA) is tokenized and embedded with the language model’s embedding layer. Those embeddings are concatenated with the embeddings built from the input vector to construct an input sequence for the language model.

For the sake of simplicity, we train the model generatively for all tasks [7, 17]. The model generates a caption, a free-form question answer, or a class name for the tasks of captioning, VQA, and visual entailment respectively. During training, the language model and linear layer are fine-

tuned, but the text encoder is kept frozen to ensure the correspondence between text and image vectors learned during pre-training is preserved.

2.3. Modality Gap

In practice, text and image vectors from contrastive models can still be far apart, a phenomenon known as the modality gap [35]. For example, on COCO captions [5] the average cosine similarity between an image and a paired caption is only 0.26, while the average similarity between two unrelated captions is 0.35. Figure 3a shows this gap causes image and text vectors to fall into separate clusters in the vector space. The root cause is that the cross-entropy loss used by contrastive models only requires paired image and text vectors to be close *relative* to random image and text pairs, which does not necessarily mean they are close in absolute terms, see Liang et al. [35] for more discussion.

We find a simple and effective solution – adding Gaussian noise that is drawn from a standard normal distribution and then scaled by a hyper-parameter w , to the text vectors during training. Intuitively, this noise requires the model to be more robust to minor changes or variations to the input vectors, and thus be better prepared for the shift caused by switching from text to image vectors. Figure 3b visually shows that even noise with a scale of $w = 0.1$ leads to better overlapping vector spaces. Adding noise also has a regularizing effect of training the model to predict similar outputs when the semantics of the input are slightly altered. After adding the noise we re-normalize the vector so it remains of unit length to match the image vectors that will be used during evaluation. We study the modality gap and other possible approaches in more detail in Section 4.

3. Experiments

We report results on three V&L tasks, captioning, visual entailment, and visual question answering, when training models using only text data, as well as on captioning when using data generated from a language model.

Model	B-4	M	C	S
ClipCap [42]	33.5	27.5	113.1	21.1
CLOSE w/Images	34.4	27.8	113.2	20.4
ZeroCap [57]	2.6	11.5	14.6	5.5
Socratic Models ZS* [76]	6.9	15.0	44.5	10.1
MAGIC [54]	12.9	17.4	49.3	11.3
CLOSE w/o Noise (Single)	4.2	12.2	16.4	6.5
CLOSE w/o Noise (Mult.)	21.9	20.6	68.7	13.5
CLOSE (Single)	28.6	25.2	95.4	18.1
CLOSE (Mult.)	29.5	25.6	98.4	18.3

* results on sample of 100 captions

Table 1. Results on the caption test set, (Single) indicates the single-caption setting and (Multi.) the multiple captioning setting.

3.1. Setup

We use $T5_{base}$ [48], $CLIP_{ViT-L/14}$, and a fixed set of hyper-parameters for all tasks (see supp. for details and ablations), but tune the noise level in the adapter on the validation set for each task.

Our primary point of comparison is with the same model trained on images, in which case the images are encoded with the image encoder in the same manner as done during testing. This model does not experience domain shift, so we view it as an upper bound for our method. We also compare to training without Gaussian noise, and with zero-shot methods from prior work when available, since they also do not require visual training data.

3.2. Image Captioning

We consider two image captioning settings. First, *multiple-caption*, when multiple captions about the same scene are available. In this case, we find it beneficial to use one of those captions as input and a different caption as the target during training. Multiple captions might not be always available (e.g., our text-only stylistic captioning datasets), so we also consider a *single-caption* setting where captions are not grouped, in which case we use the same caption as the input and target output.

We perform experiments on the COCO Captioning [5] dataset using the Karpathy split [25], and report BLEU-4 (B-4) [45], CIDEr (C) [62], METEOR (M) [9] and SPICE (S) [9] on the test set. We train text-only models by training on just the captions in the training data. We treat all captions per an image as a group for the multiple-caption setting and use each caption individually in the single-caption setting.

Results are shown in Table 1. We validate our model by comparing it to another CLIP-based captioning model, ClipCap [42], and find it performs similarly when trained on images. CLOSE achieves 98.4 and 95.4 CIDEr in the multiple and single caption setting, showing high captioning competency despite not using images. Our approach is

Model	Val	Test
CLOSE w/Images	77.0	77.7
CLIP Classifier [53]	67.2	66.6
CLOSE w/o Noise	68.7	68.2
CLOSE	75.9	75.9

Table 2. Results on the visual entailment test and validation set.

Model	Yes/No	Num.	Other	All
CLOSE w/Images	80.4	48.4	64.1	67.9
CLOSE w/o Noise	76.8	36.8	53.9	59.8
CLOSE	78.2	46.0	59.5	64.3

Table 3. Results on the VQA-E validation set.

Model	Yes/No	Num.	Other	All
CLOSE w/Images	83.2	44.8	54.9	65.4
TAP- $C_{ViT-B/16}$ [53]	71.4	20.9	18.6	38.7
CLOSE w/o Noise	78.6	40.6	49.0	60.2
CLOSE	79.4	43.4	51.1	61.9

Table 4. Results on the VQA 2.0 test-dev set.

substantially better than zero-shot methods [57, 76], which can be partly attributed to the fact that zero-shot approaches have no mechanism of learning the desired style of the target caption, whereas our approach can learn it from the text-only training data. Removing Gaussian noise results in significantly worse performance, particularly for the single caption setting, likely because simply repeating the input string becomes too easy of a task.

3.3. Visual Entailment

Visual entailment requires determining whether a premise image either entails, contradicts, or is neutral with respect to a hypothesis sentence. During training, a text premise is used instead of an image. The hypothesis sentence is always text and is encoded with T5. We train on SNLI [40] (a language only dataset) and evaluate on SNLI-VE [70] (a vision and language dataset).

The results are shown in Table 2. Despite not using images, CLOSE achieves similar performance to the image model. Song et al. [53] also experiment with this as a cross-modal transfer learning task, but we find adding Gaussian noise allows us to surpass their result.

3.4. VQA

To train a VQA model we use data that contains a short sentence describing a scene (encoded with the text encoder), a question (encoded with T5), and a target answer. We consider two sources of training data, first we pair COCO

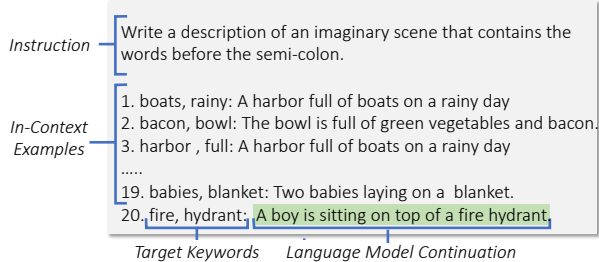


Figure 4. Prompt used to generate a synthetic caption from a language model. The language model’s continuation (highlighted text) is used as a synthetic caption.

Model	B-4	M	C	S
MAGIC [54]	12.9	17.4	49.3	11.3
CLOSE w/COCO	29.5	25.6	98.4	18.3
CLOSE w/GPT-J RNG	19.6	20.9	63.2	13.8
CLOSE w/GPT-J Unigram	23.2	22.2	78.9	15.6
CLOSE w/OpenAI Curie	18.5	21.2	69.0	14.9

Table 5. Results on the COCO validation set when training on synthetic captions. Test results from Table 1 are included above the dashed line for reference.

captions with questions about the same image from VQA 2.0 [14]. However, in this dataset, the questions might ask about details of the image not included in the caption, and thus cannot be answered by the text-only model. Hence we also train and evaluate on VQA-E [31] which contains a subset of the VQA 2.0 questions paired with COCO captions that have been verified to contain the answer.

These training sets have significantly different question distributions due to the filtering done in VQA-E (e.g., VQA-E does not include questions where the answer is “zero”), so we evaluate models either on the VQA 2.0 test-dev set or the VQA-E validation set¹ depending on what train set was used. Results for VQA-E are shown in Table 3, and for VQA 2.0 in Table 4. We compare to TAP- $C_{ViT-B/16}$ [53], a CLIP-based zero-shot approach.

For VQA-E, we observe only a 3.5 point drop relative to image training while surpassing the baselines. The gap is more significant on VQA 2.0, which we attribute to the sometimes poor alignment between the captions and questions, although our method is still within 5 points of the model trained on images.

3.5. Training with Data from a Language Model

Next, we use CLOSE to train a captioning model on synthetic data generated by a language model. We first construct a prompt that includes a natural language instruction and some example captions following an in-context learn-

¹VQA-E does not have a test set

ing approach [3], shown in Figure 4. To generate a diverse set of captions, we prefix each caption with two keywords that occur in that caption and end the prompt with two new keywords to be used in the caption to be generated. Then diverse captions can be constructed by changing the ending keyword pair. To reduce the chance of caption style affecting the quantitative evaluation, we take steps to better match the style of the COCO captions, although in settings where the precise style is of less importance this would not be required. We generate 100k examples from three generation methods (see the supplementary for additional details):

GPT-J RNG. Examples are generated using a 6 billion parameter open source language model, GPT-J [64], with 50 in-context examples. Keywords are sampled uniformly at random from keywords in the COCO training data.

GPT-J Unigram. Keywords are instead sampled to match the unigram distribution of COCO captions.

Curie Unigram. Generations are from OpenAI Curie² with 20 examples and unigram-matching.

Results on COCO are shown in Table 5. Our best result achieves 78 CIDEr. Inspection shows that, even with our keyword sampling approach, many errors are still caused by style issues, and that style also explains the reduced performance of the Curie model. For example, the synthetic captions from the Curie model are 23x more likely than the COCO and the GPT-J captions to use the word “opens” (e.g., “a living room that opens on to the balcony”), and use “cell-phone” while “cell phone” is more common in COCO captions. More details are in the supplementary. This illustrates how, when using this method, the choice of language model can have subtle effects on the style of captioning that will be learned. Despite this issue, this is still a strong result that surpasses the zero-shot methods. The method of keyword sampling is also important for style matching, we see a 15 point drop in CIDEr when using random keywords.

4. Analysis

Our approach opens up two intriguing questions: (1) Why does embedding substitution work even when text and image vectors are generally quite far apart? (2) Can methods that leverage additional data to better close the modality gap improve upon this approach? We do two analyses to answer these questions.

4.1. Sensitivity Analysis

To help answer the first question, we perform a sensitivity analysis on the input text vectors. To do this, the model is trained while adding a constant vector to the normalized text vectors and then re-normalizing, and tested on the unaltered image vectors as before. This alteration will change how the text vectors are distributed relative to the image vectors, but

²<https://beta.openai.com/docs/models/gpt-3>

Bias	Mag.	MG	Δ	Cap.	VE	VQA
none	0.0	0.26	1.00	94.4	64.3	75.9
mean	0.8	0.62	0.69	92.8	64.7	75.4
-mean	0.8	-0.10	0.85	84.3	62.0	71.8
RNG	0.2	0.25	0.98	93.5	63.9	75.3
RNG	0.5	0.24	0.89	92.5	64.2	75.3
RNG	0.8	0.20	0.78	89.3	63.7	74.8
RNG	1.0	0.18	0.71	87.2	63.8	74.2
RNG	2.0	0.11	0.45	73.7	61.4	71.3

Table 6. Text vector translation sensitivity analysis. The first three columns show the translation magnitude, the resulting modality gap on COCO, and the cosine similarity to the original vectors. The following columns show CIDEr captioning score, accuracy on VQA-E, and accuracy on visual entailment on validation sets.

will not change how the text vectors are distributed relative to one another. We show results both when using a random vector (note the same vector is used throughout all of training, it will just be selected randomly at the start of training) of different magnitudes, the mean difference of text and image vectors to represent a shift towards the image vectors, and the negation of that vector to show a shift away from the image vectors. In all cases, we continue to add Gaussian noise as before.

Results are shown in Table 6. For random vectors, we report the average of three runs with 3 different random vectors. Overall, we see only minor degradation when using random vectors until very large shifts are used, showing the model is generally insensitive to shifting the text vectors during training. Shifting the vectors towards the images (mean) can result in a slight increase in performance, and shifting the vectors away from them (-mean) results in a more significant decrease, showing the model is not completely insensitive. However it is still notable that vector substitutions works well even as the text vector’s positions are significantly randomized.

We hypothesize that this insensitivity is due to two reasons. First, most directions in the shifted feature space are predictive of the output in the same manner as before because the text vectors do not change relative positions. Second, the Gaussian noise trains the model to be insensitive to shifts in unimportant directions in the feature space, which often include the direction of the shift. This insensitivity provides part of the answer to question 1. A major source of the modality gap is a constant shift between the image and text vectors [34]. However, addressing this is not as important as one might expect because CLOSE is not highly sensitive to the absolute positioning of the text vectors.

4.2. Learned Adapter Analysis

As suggested by Figure 3c, mean shift might not be perfect at aligning the text and image vectors, so we hypoth-

Method	MG	Cap.	VE	VQA
CLOSE	0.26	94.3	75.9	64.3
+Cov. (COCO)	0.62	106.5	75.5	65.5
+Cov. (CC3M)	0.58	95.1	75.8	65.0
+Linear (COCO)	0.81	99.5	76.0	65.7
+Linear (CC3M)	0.75	81.8	75.5	64.9

Table 7. Results with adapters built with paired data. The modality gap on COCO captions, captioning CIDEr, visual entailment accuracy, and VQA-E accuracy on validation sets are shown.

esize more sophisticated adaption methods could improve performance. To learn the relationship between image and text vectors these methods require paired data, so we avoid using these approaches with CLOSE. However we still investigate them to better understand how much performance they could potentially contribute. We also study the difference between using high-quality annotated data or web data by using both COCO captions and Conceptual Captions 3 Million (CC3M) [50]. To train the adapters, for COCO we use the 30k captions from the “restval” set from the Karpathy split, which do not appear in our train or eval sets, and for CC3M we use a random sample of 100k image/text pairs. We consider two adapters:

Linear Adapter. We learn the modality shift by training a linear model to minimize the Euclidean distance between the adapted text vector and paired image vector. We continue to add Gaussian noise after applying this model.

Structured Noise. Even in principle, we do not expect there to be a perfect one-to-one mapping between text and image vectors because an image vector can be similar to many different texts that describe different parts or details of the image. This motivates us to approach the problem from the perspective of better understanding how text vectors are distributed around the related image vectors, instead of just trying to learn a simple mapping function. In the supplementary, we show the vector differences from COCO image-caption pairs do tend to follow a particular shape. To better account for this structured relationship during training, we add Gaussian noise with the mean and covariance of the differences between paired image and text vectors in the auxiliary corpus to the text vectors during training. This noise is expected to better simulate the text-image shift that will occur during evaluation.

Results are shown in Table 7. We observe large improvements on captioning, modest improvements on VQA, and no improvement on visual entailment using the adapter from COCO, with the structured noise approach being significantly better on captioning and slightly worse on the other tasks. The CC3M adapter still achieves mild gains, although it is less effective. This shows the training data used for the adapter is important, a point that can be qualitatively observed in Figure 3c and Figure 3e.

Egocentric Captions



Uplifting Captions



Harry Potter Captions



Reviews Captions



Figure 5. Examples of stylistic captions produced by captioning models trained on text with CLOSE, and then applied zero-shot to images.

5. Stylistic Captioning

We demonstrate an application of our method by applying it to the task of constructing captions with specific writing styles. Our general approach is to gather text-only training data that exemplifies the style we want the model to use, train on them as if they were text captions as done in Section 3.2, and then apply the model to images. To show that a diverse range of natural language data sources can be used to learn different styles we show four captioning styles, each of which uses a different method of collecting training data.

Ego-Centric. Section 3.5 shows that our model can be trained using data generated by a language model. Now we demonstrate an application of that approach by using the language model to generate captions in an ego-centric style. We use the same prompt format as before (Figure 4), only now with 20 examples of manually authored captions writ-

ten from a first-person perspective. We again sample keywords randomly from those found in COCO training captions to generate diverse prompts and generate 20k captions using OpenAI’s GPT-3 model. We apply this model to COCO validation images, shown in the top row of Figure 5, and observe it learns to use a variety of first-person language while accurately describing the image.

Uplifting. We use a publicly available dataset [12] to collect 6k examples of uplifting captions. Results are shown in the second row in Figure 5, where we observe the model adds warm and optimistic details to its captions.

Character-Based. Next, we target character-based captions that use proper nouns and describe images as if they are from a story. Using proper nouns would be a significant hurdle for many existing systems due to the lack of image/name paired data in existing datasets. However, it is easy for our approach since CLIP already recognizes the

names of many famous people [47], and the text training data approach teaches the model how to leverage that information when writing captions. We pick 33 Harry Potter characters to train the model on. The number of characters is limited to reduce the need to generate a very large training set. Then excerpts from the Harry Potter books, or related fan fictions, were manually collected and used as prompts to GPT-3 to create 13k text captions. Results on relevant photos are shown in the third row of Figure 5. The model uses the correct names and image content, while sometimes making up plausible events that could give additional context to the image as if it was a scene in a book or a movie.

Reviews. We train a model to write captions like a customer writing a review. For training data, we gather publicly-available Amazon product reviews³ and select positive reviews that are a maximum of 40 tokens long. As shown in Figure 5 bottom row, the captions use a variety of language to write positive reviews of the items in the photos.

6. Related Work

Cross-Modal Transfer Learning. Transfer learning has typically focused on transferring skills from one modality to the same modality. CROMA is an exception and uses a modality-invariant feature space to achieve transfer similar to our work, however, it is limited to classification tasks and is few-shot rather than zero-shot [34]. Pre-trained language models have been shown to learn skills that can transfer to new modalities [37], however, this will be ineffective for task-specific skills such as a desired captioning style or learning the space of output labels. Several multi-modal/multi-task models have learned many tasks in different modalities simultaneously [23, 33, 36, 66] and could thus potentially transfer skills between them, with HighMMT in particular showing positive results [33]. Our work studies the more challenging zero-shot setting (meaning no training data in the target modality is available), and therefore requires all the needed skills to be learned from a modality different than the one used in evaluation. Concurrently with our work, Nukrai et al. [44] propose a similar text-training approach with CLIP and apply it to stylistic captioning, although we consider additional tasks and analyses.

Using Contrastive Models. Many vision and language contrastive models have been constructed, including CLIP [47], ALIGN [22], UniCL [72] and OpenCLIP [21], and recent multi-modal models that contain a contrastive training component [29, 74, 75]. Typically these models are used either zero-shot, which is effective for image classification but challenging for more complex tasks like captioning or visual entailment [53, 57, 76], or as feature extractors for down-stream tasks [11, 15, 26, 39, 46, 51, 68, 77]. Our

work offers a compromise between those two approaches by allowing models to be trained with only textual data, which substantially improves upon zero-shot performance without requiring annotated images. The only prior work using a text-only approach that we are aware of has been in visual entailment [53], however, their study did not consider other applications and uses a relatively shallow model without the use of noise or other methods that address the modality gap.

Domain Invariant Representations. Using domain-invariant features to achieve out-of-domain generalization has a long history in transfer learning. Work in this area has shown such features can be built from multi-domain training data [16, 65], small amounts of labelled data in the target domain [8, 59], and unsupervised data [55, 67]. Methods include using adversarial learning to remove domain-dependent features [13, 32, 58], using maximum mean discrepancy to ensure features are distributed similarly across multiple domains [2, 28] and various data augmentation approaches to prevent models from learning domain-dependent features [49, 63, 79, 80]. The effectiveness of Gaussian noise in making models robust to domain shifts in these features has also been observed in image classification [30]. While we also use domain-invariant features, the domain shift we study is more extreme than what is typically studied due to the change in modalities, and we show large-scale contrastive models can be an effective source of invariant features if used correctly.

Stylistic Captioning. Stylistic captioning models can be built by authoring captions of the desired style [12, 18, 41, 52] and applying standard captioning methods. However, since creating such annotations is expensive, many stylistic captioning methods attempt to additionally transfer from captions with other styles by pre-training or multi-tasking [41, 43, 73]. Other stylistic captioning models have combined unstylized captioning data with text data in the desired style through methods such as adversarial learning [4], multi-tasking with language modelling [12], or factoring caption writing into style and context components so that the style component can be learned from the text [12, 78]. Most similar to our work, Tan et al. [56] train a model to generate text from either images or text using a shared encoding space and learned style embeddings. Unlike these methods, our approach does not require the use of any paired image/caption data.

7. Conclusion

We have shown that the multi-modal semantic vector space learned by contrastive models can be used for cross-modal generalization through CLOSE, demonstrated an application to stylistic captioning, and studied sensitivity and trained adapters. As more powerful contrastive models that span more modalities are trained, we expect CLOSE to gain more applications and yield better models.

³<https://www.kaggle.com/datasets/bittlingmayer/amazonreviews>

References

- [1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022. 1
- [2] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 8
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. 2, 5
- [4] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 521–530, 2017. 8
- [5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015. 3, 4
- [6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *ArXiv*, abs/1909.11740, 2019. 3
- [7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. *ArXiv*, abs/2102.02779, 2021. 3
- [8] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009. 8
- [9] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 4
- [10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022. 1
- [11] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 8
- [12] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 955–964, 2017. 7, 8
- [13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 8
- [14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5
- [15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 8
- [16] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 8
- [17] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems. *ArXiv*, abs/2104.00743, 2021. 3
- [18] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhat-tacharya. Captioning images taken by people who are blind. In *ECCV*, 2020. 8
- [19] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 1
- [20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 13
- [21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Han-naneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Open-clip, July 2021. 3, 8, 13
- [22] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 3, 8
- [23] Lukasz Kaiser, Aidan N. Gomez, Noam M. Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *ArXiv*, abs/1706.05137, 2017. 8
- [24] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models. In *ECCV*, 2022. 3
- [25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 4
- [26] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022. 8

- [27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13
- [28] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 8
- [29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 8
- [30] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021. 8
- [31] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567, 2018. 5
- [32] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 8
- [33] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Shengtong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. Highmmt: Towards modality and task generalization for high-modality representation learning. *ArXiv*, abs/2203.01311, 2022. 8
- [34] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 6, 8
- [35] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *ArXiv*, abs/2203.02053, 2022. 1, 3
- [36] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Motlaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022. 8
- [37] Kevin Lu, Aditya Grover, P. Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *ArXiv*, abs/2103.05247, 2021. 8
- [38] Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. *arXiv preprint arXiv:2112.08726*, 2021. 14
- [39] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 8
- [40] Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK, Aug. 2008. Coling 2008 Organizing Committee. 4
- [41] A. Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. *ArXiv*, abs/1510.01431, 2016. 8
- [42] Ron Mokady, Amir Hertz, and Amit H Bermano. Clip-cap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 4
- [43] Omid Mohamad Nezami, Mark Dras, Stephen Wan, and Cécile Paris. Senti-attend: Image captioning using sentiment and attention. *ArXiv*, abs/1811.09789, 2018. 8
- [44] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 8
- [45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 4
- [46] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021. 8
- [47] 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. 1, 3, 8
- [48] Colin Raffel, Noam M. 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. *ArXiv*, abs/1910.10683, 2020. 4
- [49] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 8
- [50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6
- [51] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 8
- [52] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12508–12518, 2019. 8
- [53] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on

- vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022. 4, 5, 8
- [54] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 4, 5
- [55] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 8
- [56] Yutong Tan, Zheng Lin, Peng Fu, Mingyu Zheng, Lanrui Wang, Yanan Cao, and Weiping Wang. Detach and attach: Stylized image captioning without paired stylized dataset. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 8
- [57] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*, 2021. 4, 8
- [58] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 8
- [59] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 8
- [60] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 3
- [61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 3
- [62] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 4
- [63] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 8
- [64] Ben Wang. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021. 5
- [65] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 8
- [66] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 8
- [67] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020. 8
- [68] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 8
- [69] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 1
- [70] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 4
- [71] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 1
- [72] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 8
- [73] Quanzeng You, Hailin Jin, and Jiebo Luo. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *ArXiv*, abs/1801.10121, 2018. 8
- [74] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 8
- [75] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 8
- [76] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aavek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv*, 2022. 4, 8
- [77] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 8
- [78] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. Memcap: Memorizing style knowledge for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 2020. 8
- [79] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 8

- [80] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 8

Appendix

A. Hyperparameters

For all tasks, we fine-tune our model with the Adam optimizer [27] with a linear decaying learning rate starting at $3e-4$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, batch size of 128, and train for 8 epochs. We use beam search with a beam size of 5 for evaluations. We use a noise level of 0.04 for VQA, 0.08 for visual entailment, 0.14 for captioning in the single caption setting, and 0.04 for captioning in the multiple captioning setting.

B. Ablations

Cont. Model	T5 Model	Cap.	VE	VQA
ViT-L/14	base	95.4	76.1	64.3
ViT-B/32	base	91.1	75.3	61.4
RN101	base	90.0	75.4	59.8
RN50	base	90.2	75.3	60.4
RN50×4	base	92.0	75.3	61.5
RN50×16	base	93.4	74.4	62.5
RN50×64	base	96.1	75.8	64.2
OpenCLIP [21]	base	99.2	76.3	65.1
ViT-L/14	small	94.4	74.9	59.9
ViT-L/14	large	93.9	75.1	65.2

Table 8. Ablations with different T5 and contrastive models. The first column indicates which CLIP model was used, with OpenCLIP indicating we use the ViT-L/14 OpenCLIP model trained on Laion 400m [21]. The last three columns show CIDEr on COCO captioning in the single caption setting, accuracy on visual entailment, and overall accuracy on VQA-E on the validation sets.

Ablation results using different contrastive or T5 models are shown in Table 8. We find the optimal noise level for these models generally does not change as these components are ablated, so we use the same noise levels as our main results for all these experiments.

We observe a consistent decrease in performance when using CLIP versions other than ViT-L/14, with only RN50×64 being comparable, showing that our method gains effectiveness as the contrastive model becomes more powerful. We observe much less dependence on the size of the T5 model, with the large model increasing performance on VQA but not on the other tasks. The OpenCLIP model is generally more effective and boosts our captioning results to nearly 100 CIDEr, showing that switching to more recent contrastive models is an easy way to boost the performance of our method.



Word	Image	Curie Model	COCO Model
pictured (100x)		a sandwich is pictured on a white background. CIDEr: 0.76	a sandwich is sitting on a white plate. CIDEr: 1.29
lays (100x)		a cat lays on a computer keyboard. CIDEr: 0.43	a cat is laying on a laptop computer. CIDEr: 1.94
cityscape (54x)		a clock with a cityscape in the background. CIDEr: 0.44	a clock on the side of a tall building. CIDEr: 1.95
person's (13x)		a tennis racquet is seen in a person's hand. CIDEr: 0.62	a close up of a person with a tennis racket CIDEr: 1.12
sunny (3.5x)		a sunny day with people flying kites. CIDEr: 0.09	a number of people on a beach with a kite CIDEr: 0.98

Figure 6. Examples of words that are over-produced by the captioning model trained on the OpenAI Curie synthetic captions relative to the model trained on the COCO captions. The first column shows the word and how much more common it is across captions generated for images in the COCO validation set. Remaining columns shows an example image and a caption from both models with the CIDEr score computed using human-annotated captions.

Model	Individual	Any
OpenAI Curie	58.8	85.0
GPT-J	42.7	81.9

Table 9. How often generated captions contain the target keywords when generating synthetic captions using different language models. The second column shows the success rate for individual generations, and the third column shows how often any caption in the 5 captions generated per a prompt contain both keywords.

C. Generating Synthetic Captions using Language Models

In this section, we give more details about how we generate captions using language models and the results from Section 3.5. When generating captions, we use nucleus sampling [20] at $p = 0.95$ and a temperature of 1, which we find generally improves results. It is not uncommon for the caption to fail to contain both input keywords, so we sample 5 captions for each prompt and then select a caption containing the keywords if one exists, and select one randomly otherwise. The in-context example captions are prefixed by randomly chosen words that exist within that caption (ex-

cluding stop words), and we use randomly selected captions from COCO training captions as the examples. During sampling, we randomly shuffle both the order of the in-context examples and what keywords are used as prefixes for those examples to improve the diversity of the outputs. If doing unigram sampling, we keep track of the distribution of words found in the captions generated so far, and sample new keywords in proportion to how under-represented they are, while never sampling over-represented words.

Statistics for how often the input keywords are correctly included in the caption are shown in Table 9. The success rate is less than 60%, although selecting from 5 generations brings the success rate up considerably. GPT-J is worse than OpenAI Curie, but sampling extra captions helps make up for this deficiency. Future work could integrate a constrained beam search method to address this difficulty [38].

We find that about 10% of GPT-J captions are not coherent or do not describe a visual scene, while these kinds of captions almost never occur with OpenAI Curie. Overall, for GPT-J, producing 100k captions took about 50 GPU hours using a NVIDIA RTX A6000. For OpenAI Curie, each generation requires approximately 500 tokens per a query, so the total cost was about 100\$⁴. Both methods are far cheaper than annotating data.

As discussed, we observe stylistic differences occur between models trained on synthetic captions and models trained on COCO captions. A particular issue is that, while unigram sampling prevents words becoming under-represented, it still allows some words to become over-represented if the language model has a natural tendency to generate them. Figure 6 contains some examples where the model trained on OpenAI Curie captions uses words like “pictured”, “lays” or “cityscape” that almost never occur in COCO captions and thus lead to low quantitative scores even when used correctly. Interestingly, we find GPT-J is not as affected by this issue, which likely stems from differences in what data the language model was trained on. Nevertheless, the captions do still correspond well to the image content, as shown by reasonably good captioning scores despite these stylistic issues, showing it is possible to learn captioning using only synthetic data.

D. The Relationship Between Image and Text Vectors

In this section, we do additional analysis on how text and image vectors are aligned. First, we do a small case study by selecting four image/caption pairs that represent two different semantic changes, and then examining how the image or text vectors change with these changes in semantics. Result are shown in Figure 7.

We observe that the text vectors change in a consis-

tent manner when the species or position of the animal is changed, while the image vectors shift in more inconsistent directions. This is likely due to minor changes in the image semantics, such as slight changes in the animal’s appearance, background, or camera position, that are not captured in the text vectors but are encoded in the image vectors. As a result, a shift in the text vectors does not correspond to a consistent shift in the image vectors, which makes perfectly aligning image and text vectors inherently challenging. For example, a linear layer could not perfectly align these four image/text vector pairs. More generally, this inconsistency reflects the fact that the image vectors capture more details than the text vectors, and thus are rarely perfectly aligned to a single text vector. This observation motivates us to add noise to the text vectors in order to simulate the fact that image vectors can only be expected to be near, but not exactly aligned with, paired text vectors even if adapters are used.

We also analyze how image and text vectors typically differ statistically. We compute the difference between the image vector and text vector for 60k image/caption pairs from COCO captions. We center these differences, and then apply PCA. The first two plots in Figure 8 show that the first PCA dimensions explain a high portion of the variance in these differences, showing that differences are often in similar directions. We also plot the Pearson correlation coefficient for the most strongly correlated features in the third plot, showing that a number of these features are tightly correlated. The fact these differences are not randomly distributed motivates our structured noise adapter method.

⁴At the current rate of 0.002\$ per 1k tokens on 11/16/2022

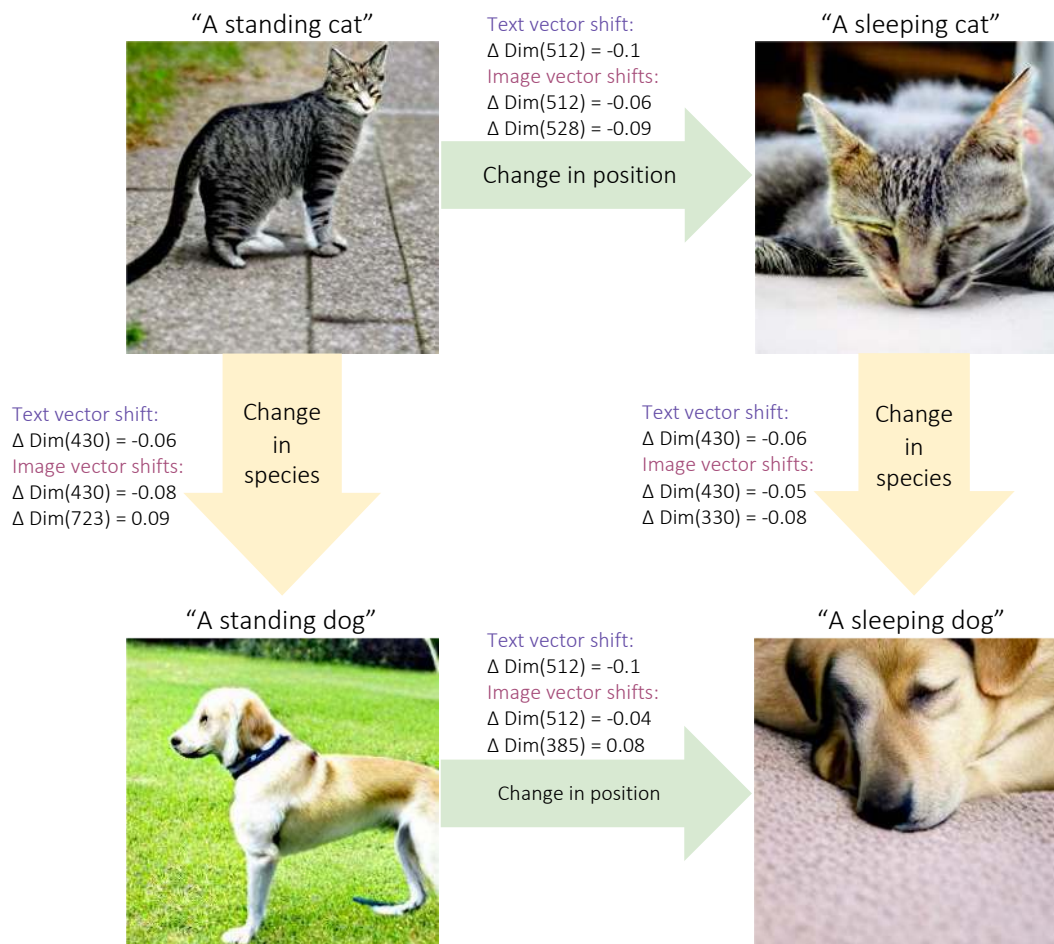


Figure 7. An example of how image/text feature vectors shift with a specific change in species (vertically) or position (horizontally). Text adjacent to each arrow shows any significant changes in the text (purple) or image (red) vector that occurred because of the shift.

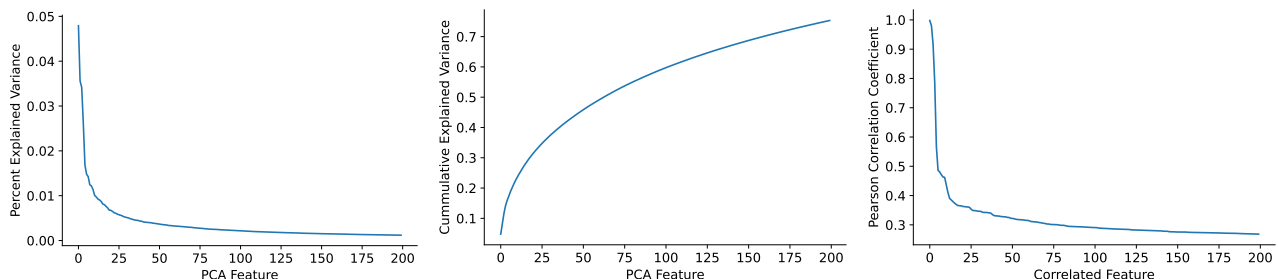


Figure 8. Plots analyzing the differences between image and text vectors for image/caption pairs in COCO captions. Only the first 200 features are shown.