Grounding Language Models to Images for Multimodal Inputs and Outputs

Jing Yu Koh [1] **Ruslan Salakhutdinov** [1] **Daniel Fried** [1]

Abstract

We propose an efficient method to ground pre-trained text-only language models to the visual domain, enabling them to process arbitrarily interleaved image-and-text data, and generate text interleaved with retrieved images. Our method leverages the abilities of language models learnt from large scale text-only pretraining, such as in-context learning and free-form text generation. We keep the language model frozen, and fine-tune input and output linear layers to enable cross-modality interactions. This allows our model to process arbitrarily interleaved image-and-text inputs, and generate free-form text interleaved with retrieved images. We achieve strong zero-shot performance on grounded tasks such as contextual image retrieval and multimodal dialogue, and showcase compelling interactive abilities. Our approach works with any off-the-shelf language model and paves the way towards an effective, general solution for leveraging pretrained language models in visually grounded settings.

1. Introduction

Trained at massive scale on large text corpora, large language models (LLMs) are able to demonstrate compelling abilities such as generating human-like dialogue and answering complex questions. While undeniably impressive, most state-of-the-art LLMs are trained on text-only data scraped from the Internet. They are not exposed to rich visual cues, and are often unable to learn concepts *grounded* in the real world. Consequently, most existing language models exhibit limitations on tasks that involve visual reasoning and grounding, and they are also incapable of producing images.

In this paper, we show that we are able to efficiently leverage

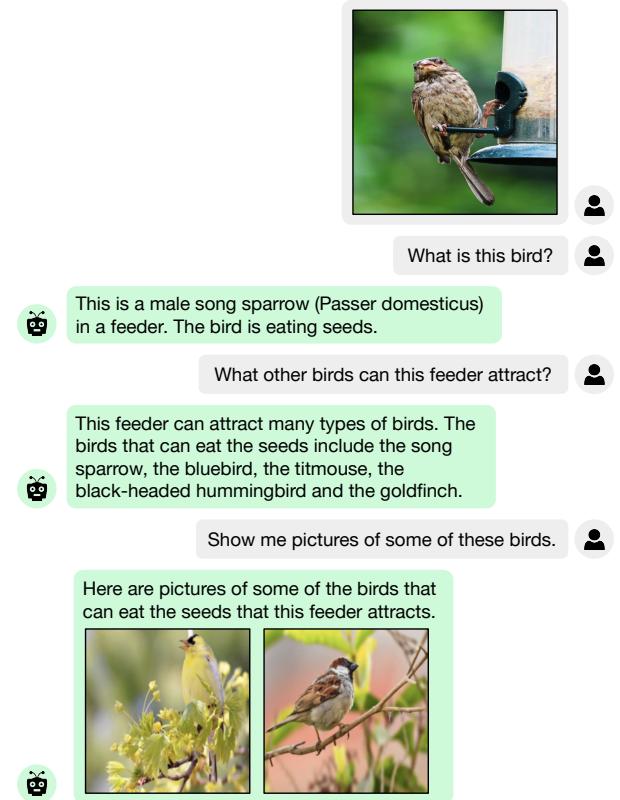[1]Carnegie Mellon University. Correspondence to: Jing Yu Koh <jingyuk@cs.cmu.edu>.

Figure 1. Our method grounds a language model to the visual domain, enabling it to process arbitrarily interleaved image-text inputs and generate coherent text outputs interleaved with relevant retrieved images. Speech bubbles in green are model generated, while grey bubbles are input prompts.

the capabilities of a frozen LLM for multimodal (image and text) input and output. Our approach equips text-only models with exciting new vision-and-language capabilities such as multimodal dialogue, generation, and contextual image retrieval from conversations (Fig. 1 and Fig. 3).

We propose a method to bootstrap a frozen language model for processing and outputting arbitrarily interleaved multimodal data. We start from a frozen pretrained LLM, and a frozen pretrained visual encoder, and train with a multi-task objective for (1) image captioning (learning to process interleaved multimodal inputs) and (2) image-text retrieval

(learning to produce interleaved multimodal outputs). For captioning, we extract visual embeddings from the visual encoder, and learn a linear mapping through the maximum-likelihood objective to map embeddings into the input space of the language model. For image-text retrieval, we train the language model to learn a new [RET] token which represents an image, and learn a linear mapping through contrastive learning (Oord et al., 2018) to map the [RET] embeddings for a caption to be close to the visual embeddings for its paired image. Most of the model is kept frozen, and we only update the weights of the linear layers and the [RET] token embedding during training. Hence, our proposed method is very computationally and memory efficient.[1] Once trained, our model exhibits several capabilities. It retains the original abilities of the text-only LLM to generate text, but also attains new multimodal dialogue and reasoning abilities. Our proposed method is model agnostic, and can be applied to ground larger or stronger LLMs released in the future. Our main contributions include:

- Proposing **F**rozen **R**etrieval **O**ver **M**ultimodal Data for **A**utoregressive **Ge**neration (FROMAGe), a model efficiently trained by visually grounding LLMs with image captioning and contrastive learning. FROMAGe learns strong few-shot multimodal abilities from image-caption pairs alone, while other models require web-scale interleaved image-text data (Alayrac et al., 2022; Aghajanyan et al., 2022).

- Demonstrating that autoregressive LLMs can perform text-to-image retrieval with greater sensitivity to input text. Our approach is more accurate on long and complex free-form text compared to existing models.

- Showing that the existing capabilities of pretrained text-only LLMs, such as in-context learning, input sensitivity, and dialogue generation, can be leveraged for visually grounded tasks. We demonstrate: (1) contextual image retrieval given sequences of interleaved images and text, (2) strong zero-shot performance on visual dialogue, and (3) improved sensitivity to discourse context for image retrieval.

Our findings pave the way towards models capable of conditioning on and generating long, coherent, multimodal sequences, and provide further insights into the abilities of pretrained text-only LLMs on visually grounded tasks. Our code and pretrained models are made publicly available[2] to encourage future work and exploration.

---

[1]Our model is trained in less than 24 hours on a single GPU.
[2]https://github.com/kohjingyu/fromage

2. Related Work

Large language models. Large language models have recently received significant attention in the machine learning and natural language processing communities, in part due to their intriguing abilities to perform in-context learning (Brown et al., 2020; Chan et al., 2022) and long-form generation (Dai et al., 2019; Tan et al., 2021; Yang et al., 2022). Most state-of-the-art models are variants of the Transformer model (Vaswani et al., 2017), with the best models achieving gains from scaling model size (Rae et al., 2021; Smith et al., 2022; Chowdhery et al., 2022), increasing pretraining data (Hoffmann et al., 2022), improving finetuning objectives (Wei et al., 2021; Tay et al., 2022), and more.

LLMs for vision-and-language. The strong performance of LLMs has also inspired work in vision-and-language research. DALL-E (Ramesh et al., 2021) proposed a Transformer based model for text-to-image generation (Reed et al., 2016) by treating images as sequences of discrete tokens. This framework was improved upon by other methods (Yu et al., 2022a; Ding et al., 2022) through model scaling, pretraining, and improved image quantization models (Esser et al., 2021; Yu et al., 2021). Flamingo (Alayrac et al., 2022) proposed a visual language model for text generation, with an impressive ability to adapt to and achieve state-of-the-art on a variety of vision-and-language tasks. Several other approaches (Wang et al., 2022; Li et al., 2022a) also propose multi-task vision-and-language pretraining approaches to improve model performance. CM3 (Aghajanyan et al., 2022) trained a causally masked model on a large HTML corpus, and showed that the model is capable of generating images and text. We differ from previous work in that our model is capable of generating coherent multimodal outputs: Flamingo is incapable of producing visual outputs, while CM3 generally produces poor visual outputs (further comparison in the appendix). In addition, FROMAGe is efficient and requires significantly less compute: it is trained in 1 GPU day (Flamingo uses 1535 TPUs for 15 days, and CM3 uses 384 GPUs for 24 days), and does not require web-scale interleaved image-text data.

Efficient adaptation of pretrained models. Lastly, our work builds upon approaches for parameter and resource efficient adaptation of pretrained models. Prefix and prompt tuning (Lester et al., 2021; Li & Liang, 2021) enable adaptation of a pretrained LLM to new settings by finetuning a small set of parameters to act as an input prefix, while keeping the rest of the model parameters frozen. Houlsby et al. (2019) proposed adapters for transferring pretrained LLMs to new language tasks. Frozen (Tsimpoukelli et al., 2021) proposed training a visual encoder to enable few-shot learning for multimodal tasks. MAGMA (Eichenberg et al., 2022) improved upon Frozen by training adapter modules for im-

proved performance on downstream tasks. ESPER (Yu et al., 2022b) uses reinforcement learning to improve zero-shot transfer and caption style transfer. Lu et al. (2022) show that language-pretrained transformers can transfer well to non-language tasks. LIMBeR (Merullo et al., 2022) analyzes pretrained vision and language models, and finds that learnt representations are functionally equivalent up to a linear transform. Our work builds upon the insights and methods from these prior works. While previous models mostly focus on generating text-only outputs, our model is capable of processing arbitrarily interleaved image-text inputs to generate coherent interleaved image-text outputs.

3. Method

Our approach integrates a language model and visual model while keeping their parameters frozen. We learn translation parameters (parameterized as linear layers) to cast images into text space, and text embeddings into visual space. Our motivation for keeping the models frozen is to leverage the capabilities of the LLM learnt from large scale pretraining. We find that this enables better generalization to zero-shot and few-shot settings (further analysis in Sec. 5.1).

3.1. Model Architecture

Language model. FROMAGe takes an autoregressive large language model p_θ, originally trained with the max-likelihood objective on text-only data, and keeps its parameters θ frozen. Given text x (e.g., an image caption), the models we use extract a sequence of input tokens $(s_1, \ldots, s_T)$ using a byte-level BPE tokenizer (Sennrich et al., 2015; Radford et al., 2019; Brown et al., 2020). The models were trained to maximize the log likelihood of the token sequence, factorized as a sum of conditional log probabilities:

$$\log p_\theta(x) = \sum_{t=1}^{t} \log p_\theta(s_t | s_1, \ldots, s_{t-1})$$

Visual model. To extract visual information from an input image y corresponding to a caption x, we use a pretrained visual backbone model which produces visual embeddings $v_\phi(y) \in \mathbb{R}^m$. The weights ϕ are kept frozen as well.

3.2. Translating Between Image-and-Text

To integrate vision and language, we learn translation parameters to map between the image and text embedding spaces. This extends a LLM for multimodal inputs and outputs.

Mapping image-to-text. We learn a linear mapping $\mathbf{W}_c \in \mathbb{R}^{m \times kd}$ which maps visual embeddings $v_\phi(y)$ from the visual model for image y as $v_\phi(y)^T \mathbf{W}_c \in \mathbb{R}^{k \times d}$ (after reshaping kd to $k \times d$). This represents a sequence of

k vectors of the same hidden dimensionality d as the text embeddings that the LLM produces for input tokens.

Mapping text-to-image. Our approach aims to retrieve images using the outputs of an autoregressive language model. A challenge with this is that autoregressive causal attention over text is strictly less expressive than the bidirectional attention typically used in previous models (Radford et al., 2021; Jia et al., 2021). In order to bootstrap strong retrieval abilities on our autoregressive LLM, we propose adding a special [RET] token to the model vocabulary and learning its embeddings (keeping all other token embeddings frozen). During training, we append [RET] to the end of input captions. This allows the model to perform an extra step of attention over all tokens in the text to produce a stronger text representation for the caption. We found that this significantly improves image-text retrieval performance (see Sec. 5.1 for analysis). This also allows our model to learn to generate [RET] at inference time (Fig. 1), seamlessly interleaving image retrieval within generated text.

Finally, to map the model's output representations to visual space, we train a linear mapping $\mathbf{W}_t \in \mathbb{R}^{p \times q}$. This maps the hidden representation of [RET] from the last hidden layer of the LLM, $h_\theta(x_i) \in \mathbb{R}^p$, into a vector space for retrieval, where q is a dimension smaller than p. Similarly, we train another linear mapping $\mathbf{W}_i \in \mathbb{R}^{m \times q}$ to map the visual embeddings $v_\phi(y_i)$ into the same retrieval space.

3.3. Training Setup

We train FROMAGe with a multi-task objective of image captioning and image-text retrieval (summarized in Fig. 2):

Image captioning. Similar to previous work (Tsimpoukelli et al., 2021; Eichenberg et al., 2022), we formulate image captioning as the task of generating text tokens conditioned on a visual prefix. The visual prefix is the output of the image-to-text mapping layer, $v_\phi(y)^T \mathbf{W}_c$, which is prepended to the caption. The log likelihood of caption x (tokenized as $(s_1, \ldots, s_T)$) conditioned on its image y is:

$$l_c(x, y) = \sum_{t=1}^{T} \log p_\theta(s_t | v_\phi(y)^T \mathbf{W}_c, \ s_1, \ldots, s_{t-1})$$

The captioning loss $\mathcal{L}_c$ is then the negative log likelihood of all samples in a batch of N image-text pairs:

$$\mathcal{L}_c = -\frac{1}{N} \sum_{i=1}^{N} l_c(x_i, y_i) \tag{1}$$

Image-text retrieval. In addition to image captioning, we train our model to retrieve images conditioned on text (and vice versa). Image-text retrieval has been used to learn
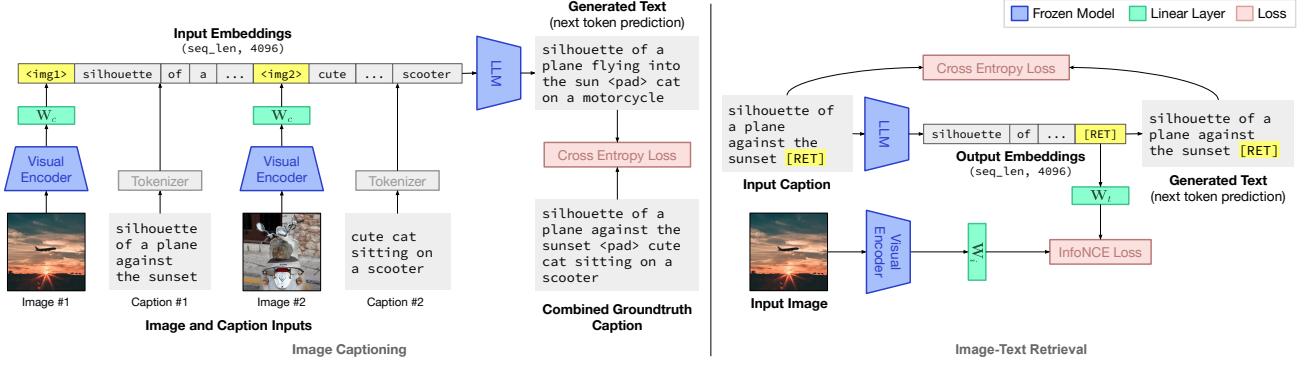
Figure 2. Overview of the FROMAGe architecture. FROMAGe is a model trained on image-text pairs for image captioning and image-text retrieval. It is capable of processing arbitrarily interleaved image and text inputs, and producing interleaved images and text as outputs.

joint visual and language representations (Jia et al., 2021; Radford et al., 2021), enabling cross-modality search from text descriptions. Underlying the approach is contrastive learning (Chopra et al., 2005) with the InfoNCE loss (Oord et al., 2018). Given a caption x and its paired image y, we extract the output of the last hidden layer of the LLM for the [RET] token, $h_\theta(x)$, and the output of the visual encoder for the image, $v_\phi(y)$. The normalized cosine similarity for the image and text embeddings can be computed with the learnt linear mappings $\mathbf{W}_t$, and $\mathbf{W}_i$ (described in Sec. 3.2):

$$\text{sim}(x, y) = \frac{(h_\theta(x)^T \mathbf{W}_t)(v_\phi(y)^T \mathbf{W}_i)^T}{\|h_\theta(x)^T \mathbf{W}_t\| \|v_\phi(y)^T \mathbf{W}_i)^T\|}$$

We minimize the InfoNCE loss for text-to-image (t2i) and image-to-text (i2t) retrieval over a batch of N text-image pairs (x_i, y_i), where each example is treated as a positive pair, and other in-batch examples as negatives:

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(x_i, y_j)/\tau)} \right) \quad (2)$$

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\text{sim}(y_i, x_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(y_i, x_j)/\tau)} \right) \quad (3)$$

Similar to previous work (Jia et al., 2021; Radford et al., 2021), τ is a learnable temperature parameter. The final training loss is a weighted sum of the captioning loss (Eq. 1) and the retrieval losses (Eq. 2 and 3):

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r (\mathcal{L}_{\text{t2i}} + \mathcal{L}_{\text{i2t}})$$

λ_c and λ_r are hyperparameters representing captioning and retrieval loss weights respectively. Since θ and ϕ are frozen, only the linear mappings $\mathbf{W}_c$, $\mathbf{W}_t$, and $\mathbf{W}_i$, and the [RET] embedding vector receive gradient updates.

3.4. Data and Implementation Details

We train on the Conceptual Captions (CC3M) dataset (Sharma et al., 2018) consisting of 3.3 mil-

lion image-text pairs.[3] To encourage the model to attend more explicitly to images, we randomly concatenate distinct examples together (with probability of 0.5 to concatenate) for the image captioning task (Fig. 2, left). We found this helpful in training the model to attend to the correct image within a sequence (detailed analysis in the appendix).

We use the publicly available OPT model (Zhang et al., 2022) with 6.7B parameters as our LLM. Past work indicates that findings at the 6.7B scale are likely to generalize to larger model sizes (Dettmers et al., 2022), and large enough to exhibit the few-shot and in-context learning abilities that we are interested in (Radford et al., 2019). For the visual model, we use a pretrained CLIP ViT-L/14 model (Radford et al., 2021) for its ability to produce strong visual representations for vision-and-language tasks (Merullo et al., 2022).
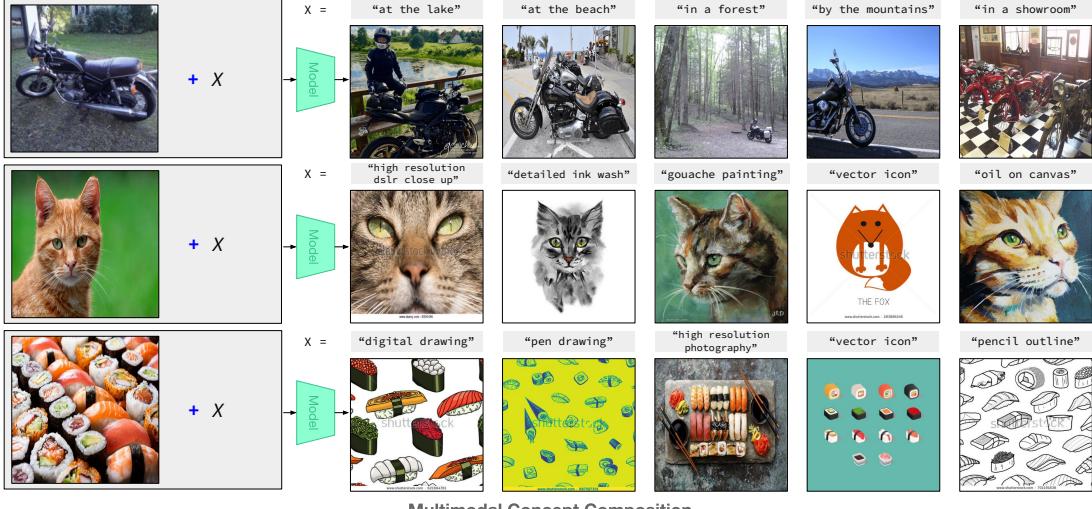
All models are implemented in PyTorch (Paszke et al., 2019) v1.12 and trained mixed-precision with bfloat16 (Abadi et al., 2016). As most of the model parameters (97.0%) are frozen, our method is memory and compute efficient: we backpropagate through the frozen LLM and visual model, but only compute gradient updates for the 3 trainable linear mappings and [RET] embedding (see Sec. 3.3). Our models are trained with a batch size of 180 for 1 epoch (18000 iterations) on 1 A6000 GPU (clock time of 24 hours). We use the Adam (Kingma & Ba, 2015) optimizer with a learning rate of 0.0003 and warmup of 100 steps. The loss weights λ_c and λ_r are set to 1 and we use a visual prefix length of $k = 1$ and retrieval embedding dimension $q = 256$, and embedding dimension $d = 4096$ (inherited from OPT-6.7B).
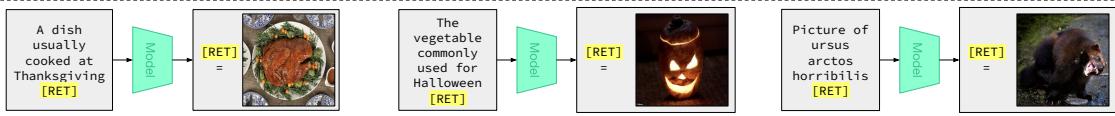
4. Experiments

The most interesting capabilities of FROMAGe emerge in situations with both image-and-text inputs and image-and-text outputs, such as multimodal dialogue (Fig. 1). As there does not exist comprehensive benchmarks for these specific tasks, we focus evaluation on image retrieval and image-

---

[3]3.1M examples remain after filtering out missing images.

Multimodal Concept Composition
FROMAGe can seamlessly composite image and text data to produce images with the desired style or content.
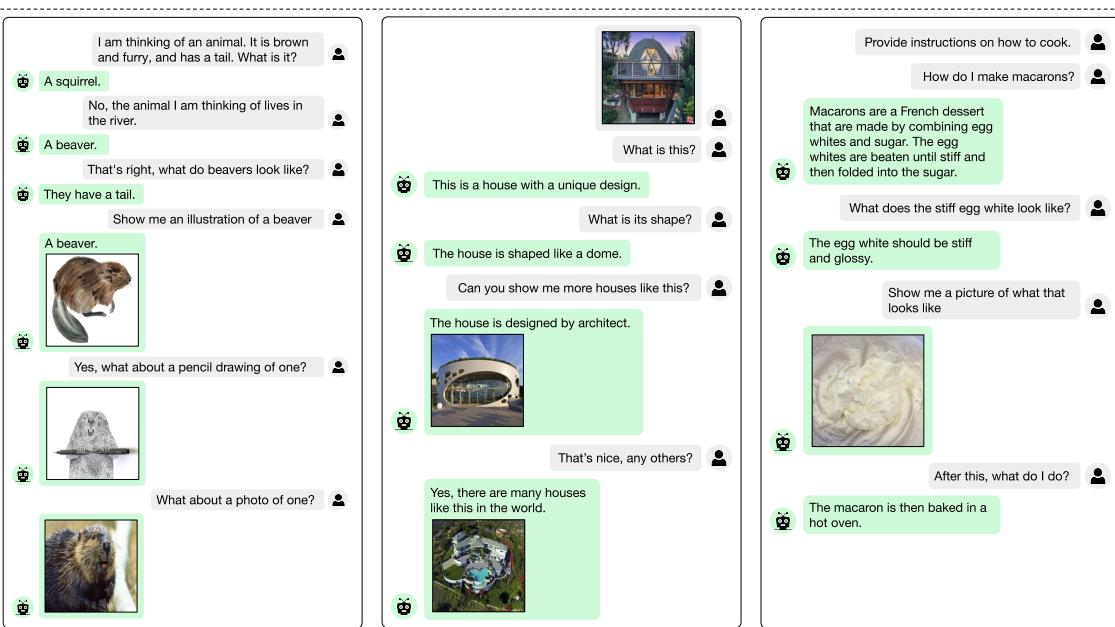
World Knowledge
Our method can draw upon knowledge learnt outside of CC3M (through large scale text pretraining of the frozen LLM) to return valid image outputs.

Visual Storytelling
FROMAGe can learn from in-context examples to generate story-like image and text outputs.

Interleaved Text-to-Image Composition
Our model can transform a sequence of text inputs into text-and-image outputs. It can do coreferencing to select the appropriate images.

Multimodal Dialogue
Our model can generate multimodal dialogue. Green bubbles indicate model generated outputs, grey bubbles indicate user provided prompts.

Figure 3. Selected examples from FROMAGe for various image-text tasks. FROMAGe is sensitive to context: it can generate multimodal dialogue, and rapidly learn in-context to perform various few shot image-text tasks. More examples are provided in the appendix.
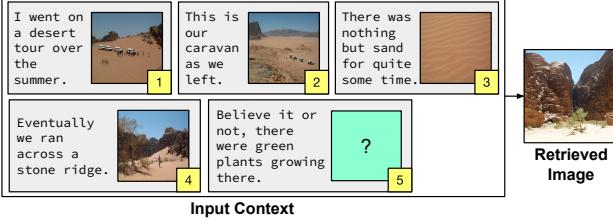
Figure 4. Contextual image retrieval conditioned on a Visual Story (Huang et al., 2016) of interleaved image-and-text inputs.

and-text generation tasks, as few prior models are capable of this. We benchmark performance on various configurations of multimodal inputs, detailed in the following sections.

4.1. Contextual Retrieval from Multimodal Inputs

Prior work on image-text retrieval (Radford et al., 2021; Jia et al., 2021) typically focuses on retrieving a single image from a single caption (and vice versa). FROMAGe is adapted from a frozen LLM, and we find that it inherits several interesting behaviors of LLMs, such as in-context learning and greater sensitivity to input context. This benefits many downstream applications, such as multimodal dialogue or image-and-text generation (examples in Fig. 3).

In order to evaluate the abilities of FROMAGe to process multimodal contextual information, we assess its performance in retrieving the appropriate image conditioned on a sequence of interleaved image-text inputs from the Visual Storytelling (VIST) dataset (Huang et al., 2016). Each example in VIST consists of 5 images and text pairs in temporal order (Fig. 4). VIST examples are of "stories", which are of a very different style compared to the image caption data FROMAGe is trained on. This allows us to evaluate our model's capability for in-context learning and zero-shot transfer. This also acts as an evaluation for discourse capabilities, as VIST contains more free-form text. We benchmark over several different experimental setups:

1. Retrieve the last image correctly given **its description**. This is similar to standard image-text retrieval.

2. Retrieve the last image given **the 5 preceding descriptions.** Image-text pairs in VIST follow a temporal order. This tests the ability of retrieval models to condition on free-form temporally dependent language.

3. Retrieve the last image given **the 5 preceding descriptions and 4 images.** This tests the ability of retrieval models to process interleaved image-and-text context.

Our results are presented in Table 1. We primarily compare against CLIP (Radford et al., 2021), as it is one of the strongest open sourced and open domain image-text retrieval models available. We report Recall@k (R@k) metrics in Tab. 1. For a single caption input, CLIP outperforms

Model	Inputs	R@1	R@5	R@10
CLIP ViT-L/14	1 caption	**11.9**	**25.5**	**32.2**
FROMAGe		9.0	21.1	28.7
CLIP ViT-L/14	5 captions	5.9	19.5	28.0
FROMAGe		**10.4**	**23.8**	**31.7**
BLIP[†]	5 captions	6.2	16.8	23.4
CLIP ViT-L/14[†]	5 captions	8.8	22.3	29.8
FROMAGe [†]	5 captions	11.6	24.7	32.8
CLIP ViT-L/14	5 captions, 4 images	2.4	21.3	34.0
FROMAGe [†]	5 captions, 4 images	**15.6**	**36.5**	**45.8**

Table 1. Recall@k on zero-shot contextual image retrieval of the last image in Visual Storytelling (Huang et al., 2016). Numbers in **bold** indicate best scores for a particular set of inputs. [†] indicates retrieval over images not previously seen in the story sequence.

our model, which we attribute to the CLIP text encoder being a bidirectional model trained specifically for image-text retrieval,[4] while our language model was trained on free-form text. However, as greater context is provided to both models, FROMAGe substantially outperforms CLIP. Given the full set of 5 captions, CLIP performance substantially deteriorates (as it appears to be unable to properly handle longer, temporally dependent sentences), with R@1 decreasing by 50.4% relative to the single caption setting. In contrast, FROMAGe is able use the additional descriptions to improve in retrieval accuracy (9.0 to 10.4 on R@1).

FROMAGe also effectively conditions on multimodal context (which previous image-text retrieval models are not explicitly trained for). When both images and text inputs are provided to the model, retrieval improves significantly, increasing by 33.6% on R@1 relative to the caption-only setting (11.6 to 15.6). Similar improvements are seen on R@5 and R@10. This is a substantial gain over the baseline CLIP model with 5 captions: we achieve a relative improvement of 77.3% on R@1 (8.8 to 15.6) when image-and-text context is provided.

We also run an experiment to test the ability of CLIP to retrieve images conditioned on multimodal context. We embed each of the images and descriptions in the input, and average their embeddings. We find that it does substantially worse than when only caption inputs are provided: it achieves a R@1 of 2.4, a significant decrease from the CLIP R@1 of 8.8 when it is provided with 5 captions. likely because it is trained to condition on image-text inputs. These results are significantly worse than that of FROMAGe under the same settings.

These results showcase the efficacy of FROMAGe as an image-text model sensitive to complex language descriptions and multimodal context (more analysis in Sec. 5.2).

---

[4]For these same reasons, CLIP is unsuitable for dialogue, and does not have few-shot and in-context learning abilities.

It is capable of parsing interleaved multimodal inputs, and strongly outperforms CLIP for longer input descriptions. As for free-form text generation, we also run human evaluations to evaluate the ability of FROMAGe to generate stories by learning in-context from VIST examples (Sec. 5.2).

4.2. Visual Dialogue

We evaluate FROMAGe on zero-shot Visual Dialog (VisDial) (Das et al., 2017). We test its ability to (1) select the correct text answer (from 100 candidates) for a question given an image and a conversation about it (image-and-text-to-text, IT2T, which is the standard VisDial task), and (2) retrieve the correct image given a conversation about it (text-to-image, T2I). Our results are summarized in Tab. 2.

For IT2T, since FROMAGe is an autoregressive generative model, we compute the perplexity of each question and answer sequence, and select the option with the lowest perplexity. FROMAGe outperforms ESPER (Yu et al., 2022b), CLIP (Radford et al., 2021), and ViLBERT (Lu et al., 2019) on R@1, improving by 20.5% relative to ESPER. FROMAGe also achieves a competitive Mean Reciprocal Recall (MRR) of 22.0 and Normalized Discounted Cumulative Gain (NDCG) of 16.5. This is substantially higher than ViL-BERT and CLIP, but worse than ESPER. We hypothesize that this is due to differences in training: ESPER uses reinforcement learning and trains on MS-COCO (from which VisDial images are derived). Flamingo (Alayrac et al., 2022) is substantially better than all other zero-shot models, which we attribute to its larger model size (80B parameters, of which 10.2B are trainable), and larger training data of multimodal webpages (43M webpages) and image-and-text data (1.8B pairs). In contrast, FROMAGe has 5M trainable parameters and is trained on CC3M (3.1M image-text pairs). Our approach may also be applied to the Flamingo model (which uses a 70B language model backbone) to enable image retrieval, which is likely to improve overall capabilities and extend it to a greater variety of tasks. On the T2I retrieval task, FROMAGe significantly outperforms prior work, achieving a 17.5% relative improvement over CLIP on R@1. ESPER and Flamingo are trained to generate text-only outputs, and are hence incapable of this task.

Our experiments demonstrate that FROMAGe achieves competitive results on zero-shot Visual Dialogue. We emphasize that unlike previous models, FROMAGe can output *interleaved* image-and-text content, and is a more general model.

4.3. Qualitative Results

We also share selected examples covering various interaction settings in Fig. 3. FROMAGe is capable of learning in-context to perform many different zero-shot and few-shot tasks. Many of the most interesting settings are those which produce interleaved images and texts as outputs, which

prior work (Tsimpoukelli et al., 2021; Alayrac et al., 2022) is incapable of, or does not generate semantically meaningful outputs for (see appendix for further comparison with CM3 (Aghajanyan et al., 2022)). Our model is capable of holding multimodal dialogue conversations — processing input images and text and responding with coherent text and image outputs. It is able to refine input images by compositing images and text concepts. FROMAGe also inherits the world knowledge of the frozen LLM, and can answer questions that require specific real world facts.

5. Analysis

We analyze various aspects of FROMAGe to determine their effects on its overall capabilities. In all experiments, models were trained on CC3M for 24 hours on a single A6000 GPU.

5.1. Ablation Experiments

We perform several ablation experiments to validate the design choices made in FROMAGe. We provided further details and results of various other ablations in the appendix.

Freezing the LLM. We find that freezing the language model is essential to retaining in-context learning and few-shot generalization abilities. When finetuned, FROMAGe performs significantly worse on VIST and VisDial. Finetuning decreases retrieval performance on VIST (R@1 with full multimodal context decreases from 12.8 to 6.2) as well as VisDial text retrieval (R@1 from 14.6 to 1.0).

Learning a dedicated retrieval token. As described in Sec. 3.2, we add a special [RET] token to represent embeddings for retrieving images from text. When the model is trained without the [RET] token, R@1 performance on VIST (with full multimodal context) is 11.3. Adding the [RET] token improves R@1 to 15.6, which is a substantial 38.1% relative gain in retrieval accuracy. We observe similar improvements across the board for other tasks.

5.2. The Effect of Context

Multimodal context helps. Since FROMAGe can process interleaved image-text data, a natural question is on the effect of image context compared to text context. To quantify these effects, we run ablations (Fig. 5, top) varying the number of captions and images provided to the model. We measure the recall of retrieving the correct image conditioned on the provided context for VIST dataset (Huang et al., 2016). Increasing context from 1 to 5 captions significantly improves the model: R@1 increases by 29% relative to the single caption case (9.0 to 11.6). However, when we provide an additional image and text example (2 captions + 1 image), we observe an even greater improvement of 42% relative to the single caption case (9.0 to 12.8). This high-

Model	Trainable Params	Finetuning Data	IT2T					T2I		
			NDCG	MRR	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT (Lu et al., 2019)	114M	3.1M	11.6	6.9	2.6	7.2	11.3	-	-	-
CLIP ViT-L/14 (Radford et al., 2021)	300M	400M	10.9	8.5	3.1	8.7	15.9	17.7	38.9	50.2
Flamingo (Alayrac et al., 2022)	10.2B	1.8B	**52.0**	-	-	-	-		Incapable	
ESPER (Yu et al., 2022b)	4M	0.5M	22.3	**25.7**	14.6	-	-		Incapable	
FROMAGe (ours)	5.5M	3.1M	16.5	22.0	**17.6**	**20.1**	**25.1**	**20.8**	**44.9**	**56.0**

Table 2. Zero-shot results on Visual Dialog (Das et al., 2017), for image-and-text-to-text (IT2T) and text-to-image (T2I) retrieval. Unlike previous methods, FROMAGe is capable of generating free-form text interleaved with image outputs through text-to-image retrieval.
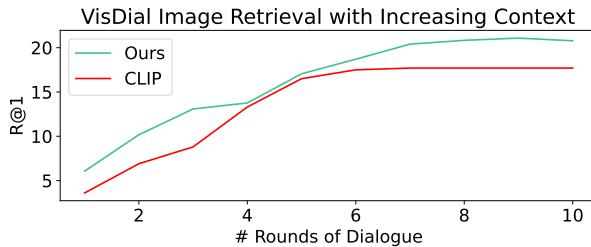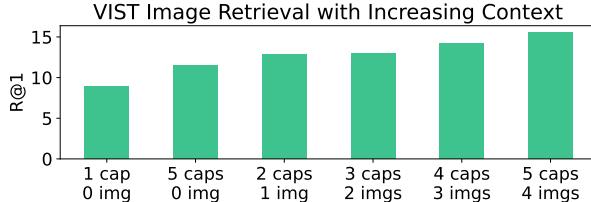
Figure 5. Increasing input context generally improves performance. Shown are results for image retrieval for VIST (Huang et al., 2016) (top) and image retrieval on VisDial (Das et al., 2017) (bottom).

Figure 6. More coherent and relevant text is generated when in-context examples are provided to FROMAGe. When multimodal context is provided, the outputs for VIST are more story-like, while the outputs for a single image input are more caption-like.

Figure 7. Human evaluations on VIST story generation. Including both images and captions improves story coherence over using just images, and improves image relevance compared to just captions.

lights the value of multimodal context: a single image can provide more information than multiple text descriptions.

More context helps. Performance also steadily improves on image retrieval for VIST as more image and caption context is provided (Fig. 5, top). The highest R@1 of 15.6 is achieved with 5 captions and 4 images (i.e., the full story context excluding the image to be retrieved), representing a 73% relative improvement over the single caption case. Similar trends are observed for image retrieval using VisDial (Das et al., 2017) dialogue rounds (Fig. 5, bottom), with performance improving as more rounds of dialogue are provided. Additionally, the results show that FROMAGe outperforms CLIP in all settings, and significantly outperforms CLIP when the full set of dialogue is provided, achieving an improvement of 17.5% relative over CLIP. These findings suggest that FROMAGe is more sensitive to context, enabling it to perform better in situations where correctly parsing long language descriptions is crucial to performance.
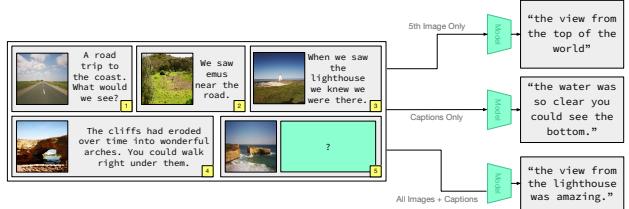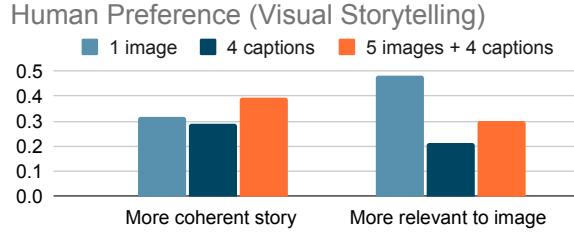
5.3. In-context Learning and Text Generation

As FROMAGe uses a frozen LLM as its backbone, it is also capable of in-context learning (Brown et al., 2020; Chan et al., 2022), where it generalizes rapidly from a few input examples. We observe this qualitatively from generating new stories for VIST, as shown in Fig. 6. When a single input image is provided, the model generally produces simple caption-like descriptions. However, when prompted with the full multimodal context (i.e., 5 images and 4 stories), the model is able to learn in-context to synthesize plausible story-like text for the 5th image (Fig. 6).

As evaluating generated text is difficult, especially for subjective outputs such as *stories*, we run human evaluations to study the effect of multimodal context on model generated stories. We request human annotators to select the output (from 3 anonymized models) which (1) forms the most coherent story when viewed in relation with the context, and

(2) is most relevant to the image. We sample 100 random examples and collect 5 independent ratings each. The results are aggregated from pairwise comparisons (details in appendix) and summarized in Fig. 7. When only the last image or the text description is provided as input, the generated stories are rated as less coherent than FROMAGe with the full multimodal context (5 images and 4 descriptions). The model generated story is also rated as significantly more relevant to the image inputs compared to the text-only setting, which highlights the ability of the model to condition on both image and text inputs. We observe that the generated output of the single image input case is rated as more relevant compared to the full multimodal context case, which we attribute to the fact that the model produces more factual (albeit less story-like) descriptions (Fig. 6). These results showcase the ability of FROMAGe to learn in-context to synthesize coherent and consistent multimodal outputs.

6. Future Work

FROMAGe is one of the first models capable of parsing image-text inputs, and producing text interleaved with retrieved images. There are several promising directions that are worth exploring in future work. Extending FROMAGe to perform novel image generation in addition to image retrieval is a natural way to improve its practical capabilities. In our qualitative experiments, we found that the ability of FROMAGe to produce relevant images was sometimes limited by its candidate retrieval set. This is often the case for prompts that are less likely to occur in natural images, such as fantastical prompts used for benchmarking text-to-image generation models (Yu et al., 2022a). On such examples, we find that FROMAGe (and other retrieval models, such as CLIP) often do not produce relevant images. Developing a model that can both generate text and novel images is an open direction which will likely require further architectural improvements. Another current limitation of FROMAGe is that it does not always generate [RET] during inference, and generally has a stronger bias to produce regular text tokens. This is likely due to its extensive pretraining on text-only data. During inference, we find that this can be somewhat alleviated by scaling the [RET] logits by a factor $1.3 - 1.5$, prompting with in-context examples, or specifically prompting the model to ask it to show images, which we found helpful in producing good qualitative results. Investigating ways to enable FROMAGe to generate [RET] more naturally is also a promising direction for future work. This may entail instruction finetuning (Wei et al., 2021) on multimodal dialogue examples, or training on explicitly interleaved image-text examples (Alayrac et al., 2022).

7. Conclusion

We propose a method to visually ground pretrained frozen language models through efficient finetuning of several linear layers. Our model, FROMAGe, is capable of producing coherent interleaved image-text outputs. We show strong zero-shot performance on a variety of tasks involving image-text inputs and outputs, and qualitatively showcase interactive abilities such as multimodal dialogue. These results demonstrate the effectiveness of our approach for bootstrapping general purpose vision-and-language models, capable of consuming and producing arbitrarily interleaved images and text. Scaling FROMAGe with larger and more capable language models, training on larger image-text datasets, and extending our approach for generation of novel images from scratch are promising directions for future work.

Acknowledgements

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

Banerjee, S. and Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *NeurIPS*, 2020.

Chan, S. C., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., McClelland, J., and Hill, F. Data distributional properties drive emergent few-shot learning in transformers. *NeurIPS*, 2022.

Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *ACL*, 2019.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. Visual dialog. In *CVPR*, 2017.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *NeurIPS*, 2022.

Ding, M., Zheng, W., Hong, W., and Tang, J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

Eichenberg, C., Black, S., Weinbach, S., Parcalabescu, L., and Frank, A. Magma–multimodal augmentation of generative models through adapter-based finetuning. *EMNLP*, 2022.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *EMNLP*, 2020.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *NeurIPS*, 2022.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *ICLR*, 2020.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

Huang, T.-H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al. Visual storytelling. In *NAACL-HLT*, 2016.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICLR*, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *ICLR*, 2015.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. *EMNLP*, 2021.

Levesque, H., Davis, E., and Morgenstern, L. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022a.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *ACL*, 2021.

Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., and Lewis, M. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019.

Lu, K., Grover, A., Abbeel, P., and Mordatch, I. Pretrained transformers as universal computation engines. *AAAI*, 2022.

Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *ICML*, 2021.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative adversarial text to image synthesis. In *ICML*, 2016.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *ACL*, 2015.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *ACL*, 2018.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

Tan, B., Yang, Z., AI-Shedivat, M., Xing, E. P., and Hu, Z. Progressive generation of long text with pretrained language models. *NAACL*, 2021.

Tay, Y., Wei, J., Chung, H. W., Tran, V. Q., So, D. R., Shakeri, S., Garcia, X., Zheng, H. S., Rao, J., Chowdhery, A., et al. Transcending scaling laws with 0.1% extra compute. *arXiv preprint arXiv:2210.11399*, 2022.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *NeurIPS*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 2017.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *ICML*, 2022.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *ICLR*, 2021.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *TMLR*, 2022.

Yang, K., Peng, N., Tian, Y., and Klein, D. Re3: Generating longer stories with recursive reprompting and revision. *EMNLP*, 2022.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *ICLR*, 2021.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022a.

Yu, Y., Chung, J., Yun, H., Hessel, J., Park, J., Lu, X., Ammanabrolu, P., Zellers, R., Bras, R. L., Kim, G., et al. Multimodal knowledge alignment with reinforcement learning. *arXiv preprint arXiv:2205.12630*, 2022b.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A. Qualitative Examples

In this appendix section, we provide more qualitative examples of FROMAGe on various settings.

Sensitivity to prompts. FROMAGe is able to tackle several examples inspired by the Winograd schema (Levesque et al., 2012). These examples contain several sentences which differ only in a single word, and contain an ambiguity resolved in different ways. Our model is capable of retrieving images correctly for different sentences, showcasing its sensitivity to even slight changes in the input prompts.
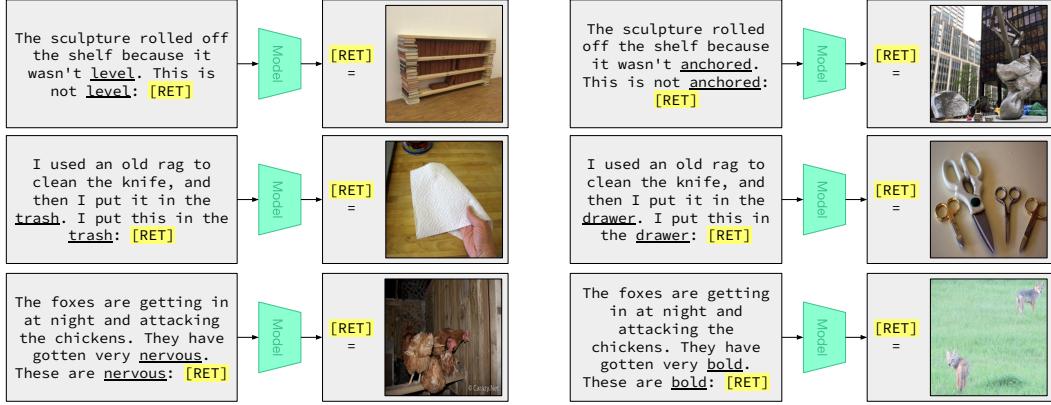
World knowledge. The FROMAGe approach involves finetuning just linear layers on the Conceptual Captions (Sharma et al., 2018) dataset, which contains image-caption data. Similar to Frozen (Tsimpoukelli et al., 2021), we find that since our frozen LLM was trained on web-scale text data, it contains knowledge about the world that it can reference for performance on multimodal tasks. For example, we show (Fig. 8) that the model knows what the weather at 0 degrees Celsius is likely to look like (snowing), that pickles are made from cucumbers, and more.

Multimodal dialogue. We also show further examples of our model on dialogue tasks. It is able to reason about input images from the user, as well as respond with semantically appropriate images in the conversation. Similar to its original LLM backbone, it can return coherent text-only outputs. It is able to tap onto its pretrained knowledge to return relevant and accurate information about the world, such as details about the water cycle (second dialogue sequence in Fig. 8) and the temperature at which water freezes (in both Fahrenheit and Celsius). This knowledge extends to the visual domain: as seen in the first dialogue sequence in Fig. 8, FROMAGe is able to understand that the photo is black and white, and likely to be taken in the 1950s.
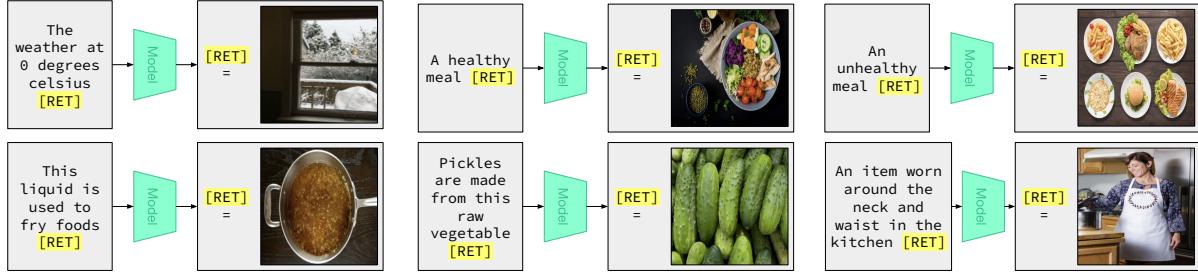
A.1. Comparison Against CM3

To the best of our knowledge, CM3 (Aghajanyan et al., 2022) is the only prior work which proposes a model capable of consuming arbitrarily interleaved image-and-text inputs and generating image-and-text outputs. CM3 trains with far larger computational resources compared to our model – they train with 384 GPUs for 24 days, while we use a single GPU for 1 day, making our method far more computationally efficient.

To benchmark the performance of the two models, we run a qualitative comparison to compare the produced images given an image-and-text story input from Visual Storytelling (VIST) (Huang et al., 2016). As FROMAGe produces images through retrieval and CM3 is a generative model, we are primarily interested in their abilities to produce semanti-
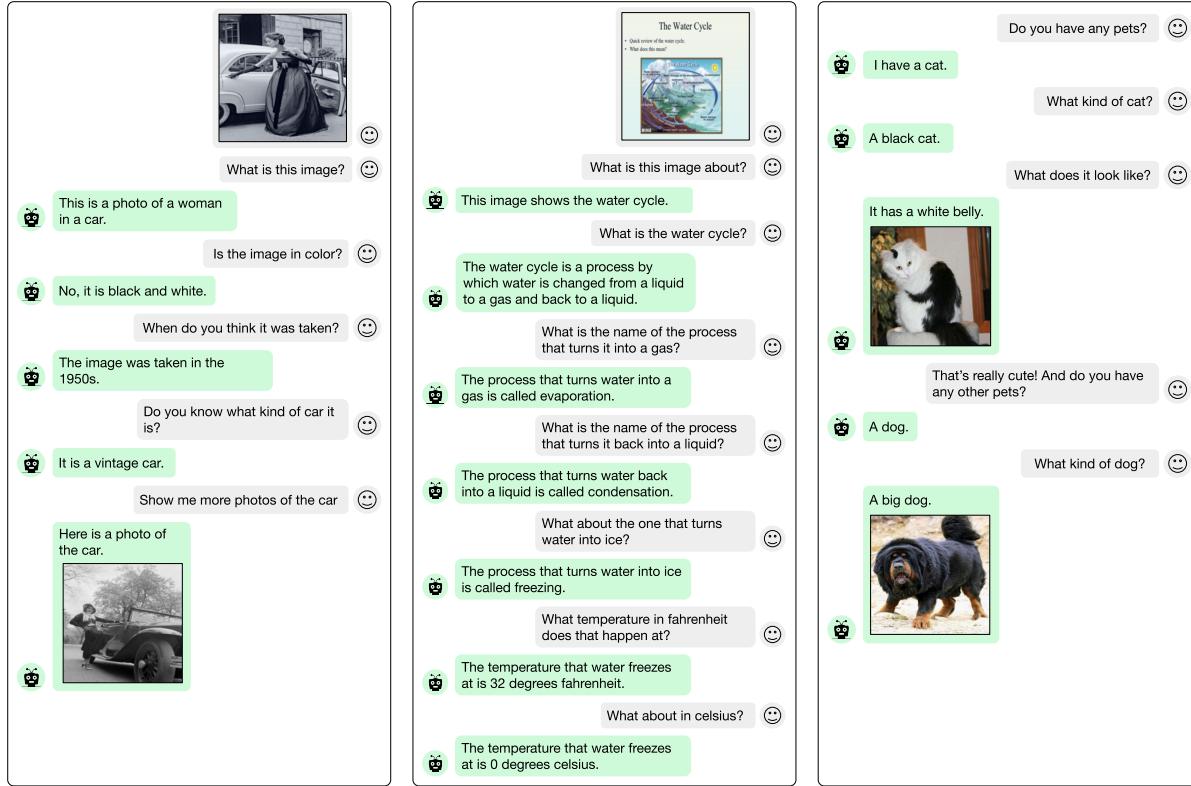
Winograd Schema Inspired Examples
Examples are pairs of sentences which differ only in one word, and contain an ambiguity resolved in opposite ways. Our model is sensitive to input language, can use world knowledge and reasoning to output correct images for certain such examples.

World Knowledge
Our method can draw upon knowledge learnt outside of CC3M (through large scale text pretraining of the frozen LLM) to return valid image outputs.

Multi-Modal Dialogue
Green bubbles indicate model generated outputs, grey bubbles indicate user provided prompts.

Figure 8. Selected examples from FROMAGe for various image-text tasks. It is capable of retrieving correct images from some examples from the Winograd schema, as well as possess world knowledge.

13
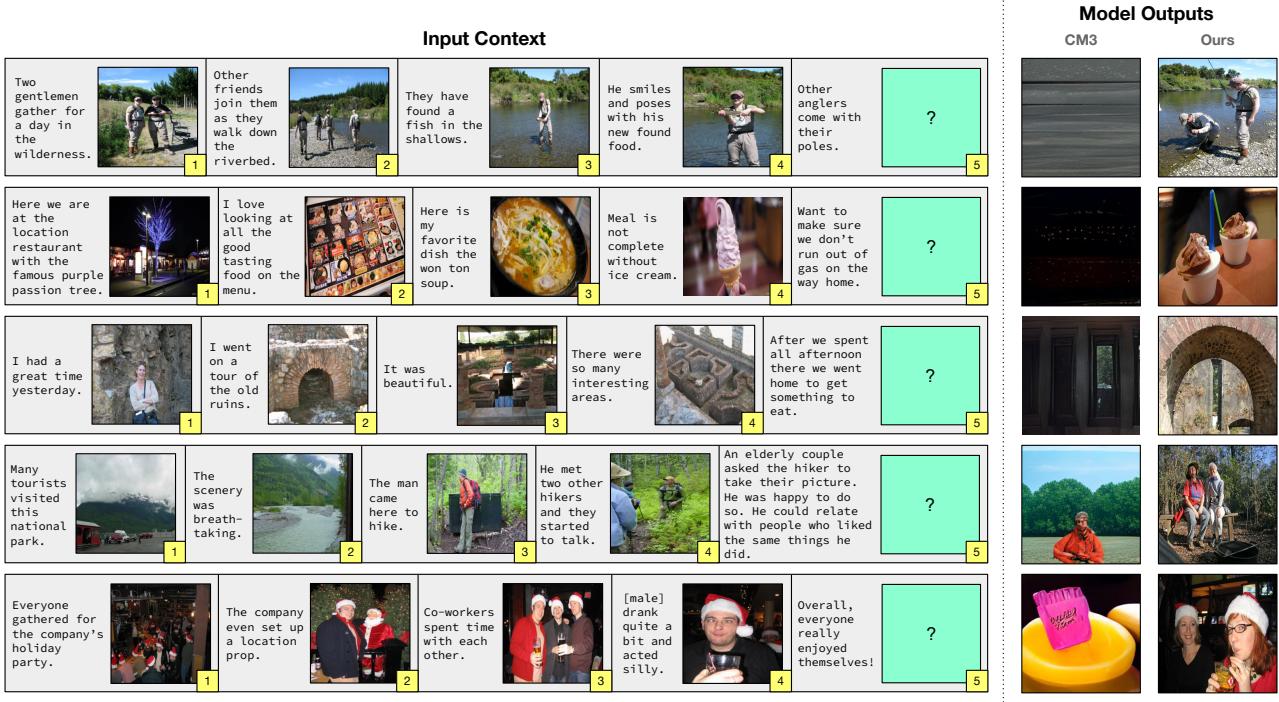
Figure 9. Comparison of our model against CM3 (Aghajanyan et al., 2022) on randomly selected examples from Visual Storytelling (VIST) (Huang et al., 2016).

cally relevant images, rather than good quality images. Several randomly selected qualitative examples are presented in Fig. 9. We observe that CM3 is unable to produce coherent outputs for most of the Visual Storytelling inputs. Most outputs produced by CM3 are not interpretable or relevant to the story input. In contrast, the outputs from FROMAGe are relevant to the inputs, and a few (e.g., first row of the fishermen, and last row of the people in Santa hats) are capable of retrieving images that are visually and semantically coherent with the input story. We also observed that in general, FROMAGe is more sensitive to input prompts and images, while CM3 does not appear to be able to handle long input sequences as well as FROMAGe.

B. Further Analysis

B.1. Details on Freezing Ablation

We explore the effect of freezing the weights of our language model. Due to GPU memory constraints, we run this experiment with the 1.3b OPT (Zhang et al., 2022) as the LLM backbone. We compare a version where the weights are kept frozen (FROMAGe with a 1.3b LLM), and a version where the language model is allowed to be finetuned. Despite the finetuned model achieving lower loss (on both the training and validation sets of CC3M), we observe that downstream performance on VIST and VisDial significantly deteriorates. On VIST, finetuning the language model de-

creases retrieval performance on $R@1$ from 12.8 to 6.2, and on VisDial (IT2T), decreases $R@1$ from 14.6 to 1.0. These results demonstrate the importance of freezing the LLM backbone in order to retain the abilities of the LLM (in-context learning, zero-shot generalization) learnt from large scale text pretraining.

B.2. Joint Retrieval + Captioning Training

We run ablations over the different loss functions (Tab. 3). As the retrieval only model is only able to process text inputs, and the captioning model is only able to generate text outputs, we are unable to test the ablated models on VIST or VisDial. Hence, we benchmark their captioning and retrieval performance on MS-COCO (Lin et al., 2014), which tests generalization from our training data (Conceptual Captions 3M (Sharma et al., 2018)).

We find that joint training with the multi-task captioning and retrieval losses have no negative effect on performance on the individual tasks, with most metrics staying the same (with captioning scores slightly improving in the FROMAGe model), which shows that we do not suffer from optimizing our model over multiple objectives.

B.3. Image-Text Concatenation for Captioning

During training, we concatenate distinct examples sequentially for image captioning. We found that this was signifi-

Training Loss	Captioning					T2I Retrieval			I2T Retrieval		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	R@1	R@5	R@10	R@1	R@5	R@10
Captioning	0.4768	0.2899	0.1664	0.0965	0.2820	-	-	-	-	-	-
Retrieval	-	-	-	-	-	23.4	47.3	59.0	26.8	52.4	63.6
Captioning + Retrieval	0.4766	0.2932	0.1720	0.1023	0.2873	23.4	47.2	58.0	26.4	52.3	63.4

Table 3. Ablation results over different training objectives. All models are trained on CC3M (Sharma et al., 2018) and reported on the 5K validation set of MS-COCO (2017) (Lin et al., 2014). For captioning, we report BLEU (Papineni et al., 2002) and METEOR (Banerjee & Lavie, 2005) scores, and for retrieval, we report Recall@k (single captions).

cantly helpful for several downstream tasks, as it encourages our model to attend to multiple images within a sequence during training. In particular, on the VIST dataset, enabling image-text concatenation improves $R@1$ from 11.6 to 15.6 when 5 captions and 4 images are provided as input (see Sec. 4.1). On VisDial, we find that the ablated model performs similarly. This agrees with intuition, as VIST requires processing of multiple images interleaved with text (while VisDial only has a single image in its input).

These results show that random concatenation is a strong data augmentation strategy for generalization to tasks involving interleaved-image-text data. As large datasets with interleaved image-text data (such as those used in Flamingo (Alayrac et al., 2022) or CM3 (Aghajanyan et al., 2022)) are generally not available to the public, our proposed approach may be a way to leverage large open image-text datasets (Schuhmann et al., 2021) for training such multimodal models.

B.4. Image-Text Concatenation for Retrieval

As described in Sec. B.3, we concatenate distinct examples sequentially for image captioning during training. This was found to be helpful in encouraging the model to learn to attend to multiple images interleaved within an image-text sequence. We ran the same experiment for concatenating examples for both image captioning and image-text retrieval. For retrieval on concatenated examples, the model is tasked to retrieve two separate images, one for the [RET] token at the end of the first caption, and one for the one at the end of the second. An example input text for a concatenated example is:

```
silhouette of a plane against the
sunset [RET] cute cat sitting on a
scooter [RET]
```

in which case the model is expected to retrieve the appropriate images of a plane and a cat respectively. However, we find that this concatenation does not appear to have a positive effect on the downstream tasks of VIST and VisDial. On VIST, $R@1$ also slightly decreases from 15.6 to
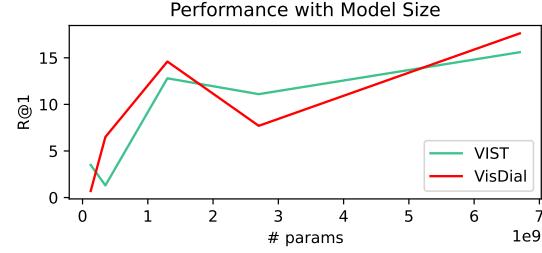

Figure 10. Performance on VIST contextual image retrieval and VisDial IT2T over different model scales. Performance generally improves as models get bigger.

14.4. We hypothesize that this is likely because these tasks (and many of our qualitative examples) do not require the model to retrieve *disparate* images – multiple image outputs are usually related (e.g., dog example in Fig. 3 of the main paper) and involve coreferencing. Concatenation of retrieval examples during training is likely to deteriorate these abilities. However, in certain multimodal applications (such as retrieval for more factual multimodal tasks, rather than dialogue), it may be possible that this retrieval concatenation strategy is still useful, and we leave exploration to future work.

B.5. Scaling Properties

FROMAGe is trained in a model-agnostic approach (Sec. 3.3), which can be applied to any pre-trained text-only LLM. We demonstrate that our approach is scalable and can benefit from larger, more expressive LLMs. We conduct experiments using the OPT model family (Zhang et al., 2022) with models of increasing parameter counts (125M, 350M, 1.3B, 2.7B, and 6.7B).

Our results, as shown in Fig. 10, indicate that performance generally improves with increasing model size on the zero-shot contextual image retrieval task for Visual Storytelling (Huang et al., 2016) and Visual Dialog (Das et al., 2017). This promising trend suggests that our framework is likely to benefit from even larger text-only models such as GPT-3 (175B) (Brown et al., 2020), Chinchilla (70B) (Hoff-

mann et al., 2022), or PaLM (540B) (Chowdhery et al., 2022). In future work, it will be interesting to test this, and trained models may learn additional interesting emergent behaviors (Wei et al., 2022).

B.6. Text Generation Results

In addition to the above evaluations, we also ran the zero-shot VQAv2 (Goyal et al., 2017). We apply the same normalization techniques from their GitHub repo[5], and prompt the model with the prefix Q: {question} A: format. On zero-shot VQA, our model achieves a score of 28.51, which is better or comparable to prior methods: a reimplementation of Frozen achieves 25.53, while running the MAGMA pretrained model (which uses 25M image-text as training data, including the VQA training set), achieves 28.35. These results show that our approach is competitive with similar methods that use parameter efficient adaptation (Frozen, MAGMA), with the added bonus that our model can perform image retrieval interleaved with text. This allows us to handle a much wider range of tasks: for example, Frozen and MAGMA cannot produce images interleaved with generated text. Our model is also significantly more efficient (1 GPU day of training) compared to prior work (MAGMA uses 32 GPUs for 1.25 days).

C. Human Evaluation Procedure

As detailed in Sec. 5.2 of the main paper, we perform human evaluations of 500 randomly selected examples to determine generated story quality given different input contexts. Users are tasked to evaluate three settings:

1. Generated outputs conditioned on the last image only.

2. Generated outputs conditioned on the preceding text descriptions only.

3. Generated outputs conditioned on the preceding images and text descriptions.

The results (described in Sec. 5.2 of the main paper), show that setting #3, which contains the full multimodal context, produces text that is rated as more coherent than both settings #1 and #2, which contain less context (which is of a single modality). We present individual ratings in Fig. 12, which summarize the head-to-head comparisons for comparing one model against another.

We observe that #3 produces descriptions that are rated as more relevant to the image than #2, but less relevant than the single image case #1. We attribute this to #1 generating more factual, caption-like outputs (as only a single image is provided as input), while #3 generates more story-like outputs (and hence are more coherent overall). Overall, these

results show the ability of FROMAGe to learn in-context to produce stories rather than caption outputs. FROMAGe is able to learn in-context and condition on both the input images and text descriptions to produce coherent story outputs which are aligned to a corresponding image (side-by-side qualitative examples in Fig. 13).

We ran evaluations on Amazon Mechanical Turk with human raters located in the US and Canada. Annotators were paid at an estimated hourly rate of 15 USD / hr. We spent a total of approximately 140 USD to collect our evaluations.

D. Current Limitations and Broader Impacts

Many existing large language models and large generative models are prone to certain types of unintended behavior. They sometimes make up facts, generate toxic and socially biased text outputs, propagate disinformation, or ignore user prompts (Gehman et al., 2020; Bommasani et al., 2021; Bender et al., 2021). When tasked to generate text, these large models also often exhibit failure modes such as neural text degeneration and generation of incoherent and repetitive text (Holtzman et al., 2020; Li et al., 2022b).

A core component of the FROMAGe model is the frozen LLM backbone, which we ground for producing text and visual outputs. Unsurprisingly, our model also inherits some of the problems that text-only LLMs possess, such as generating repetitive outputs, not following user instructions, and other common failure modes. It is also susceptible to the limitations of these language models, including the risk of producing disinformation or toxic content. The broader issues relating to text generation for our model are also likely to be addressed and alleviated by future work on better large language models. In particular, using language models that are finetuned with human feedback (Ouyang et al., 2022) or instruction finetuned (Wei et al., 2021; Chung et al., 2022) may be one direction towards improving output quality, and for reducing the risk of producing toxic and socially biased content. As FROMAGe is modular in nature, we can easily swap out our LLM backbone for better and more robust language models released in the future, enabling us to easily improve its performance on downstream applications, and reduce the risk of generating harmful content.

FROMAGe is also a model that can produce images. In this work, we produce images (interleaved within text) by retrieving from a fixed set of images from Conceptual Captions (Sharma et al., 2018). Like other image-text retrieval models (Radford et al., 2021; Jia et al., 2021), our model is susceptible to existing biases found in the training and retrieval datasets. Although image retrieval is unable to produce truly novel images from outside of the retrieval data, it also has benefits for controlling output results. Unlike image generation models which synthesize novel images

---

[5] https://github.com/GT-Vision-Lab/VQA

Figure 11. User interface shown to for human raters for performing evaluations. Raters are tasked to compare model outputs with different contexts as inputs, and rate (1) whether they form more coherent stories and (2) are more relevant to the last image. Model outputs are anonymized and shuffled.
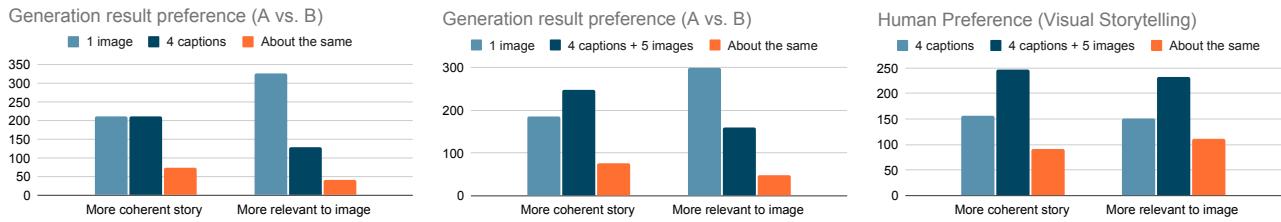

Figure 12. Head-to-head evaluations of FROMAGe with different input contexts. For each figure, human evaluators are tasked to select whether one model is more coherent than the other, and if one is more relevant to the image.
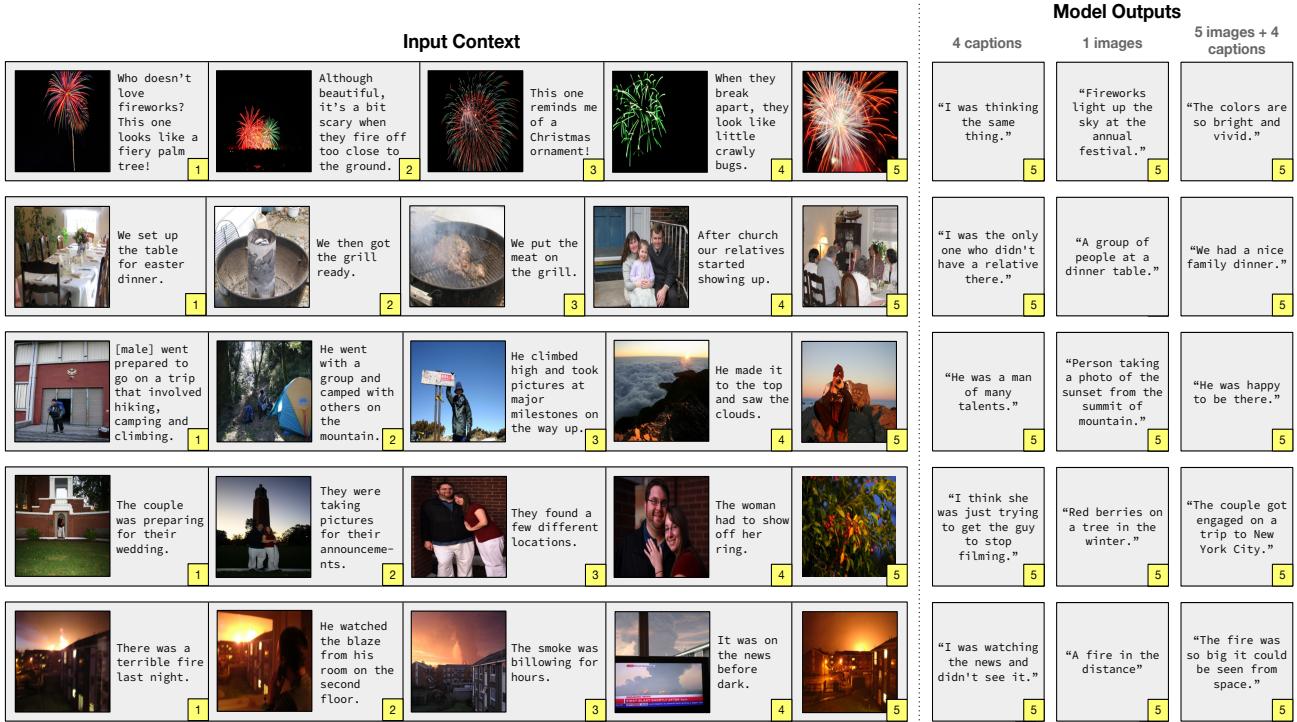

Figure 13. Examples of generated stories conditioned on different input contexts.

from scratch, a benefit of retrieving from a fixed corpus is that it allows us to explicitly control what our model can output. Retrieval enables possible mitigation strategies such as filtering for inappropriate content, such that FROMAGe and similar models would not be able to produce particular types of objectionable images. However, for deployment of such technologies (and future research on generative multimodal dialogue models), it is essential to test and analyze data used to mitigate the risk of training large multimodal models (Birhane et al., 2021). This will involve filtering of images, rigorous testing of model biases (for both image and text content), and more.